5-2021

Authorship Identification of Li Qingzhao's Anthology: A Computational Approach

Richard Tyler Simmons
Bates College, rsimmons@bates.edu

Authorship Identification of Li Qingzhao's Anthology: A Computational Approach

An Honors Thesis

Presented to

The Faculty of the Chinese Program

Bates College

In Partial Fulfillment of the requirements for the

Degree of Bachelor of Arts

By

Richard Tyler Simmons

Lewiston, Maine

5 May 2021

Table of Contents

Introduction

Knowing an author's identity is a critical element in analyzing most written works because it allows the reader to better understand the context of the writing. Imagine if people could read any document, text, novel, or poem, and know exactly who wrote it. Writers of anonymous ransom notes would never be safe and students would never have to write their names on their homework ever again. Although this sounds impossible, there are computational tools that help us to make judgments about authorship. One of these tools is machine learning authorship identification algorithms, and this technology is being used by police to read anonymous internet posts, by computer scientists to build apps, and by academics to study authorship of certain historical texts. But what exactly is authorship and why is it significant? Authorship is an important tool while analyzing any work of art (novels, poems, prose, etc.) and is defined as the "state or fact of being the writer of a book, article, or document, or the creator of a work of art". It often helps us understand the writer's thought process and message more clearly. In particular, correlating the meaning of a poem to the biographical detail of the author is one of the most dominant ways of traditional Chinese poetic exegesis. However, authorship is a tricky concept, especially those of ancient works. The names of authors and their histories can be lost or misattributed over time, leading to an uncertainty of authorship for many famous texts. This Chinese Studies thesis will investigate the authorship of Li Qingzhao 李清照 (1084–ca. 1155), the most famous Chinese female poet, and her anthology by using computational tools like natural language processing techniques. More specifically, this thesis is an exploration of Li Qingzhao's anthology and history, an investigation into authorship identification techniques, and a case study of implementing this technology on Li Qingzhao's anthology.

Li Qingzhao is widely accepted as China's greatest female poet. As a woman whose father was the student of Su Shi 苏轼 (1037-1101), one of the most prominent poet-officials in pre-modern Chinese history, Li Qingzhao's work was already well-known in the Song dynasty. She was a very prominent scholar in her times, sometimes even writing for governors in her province. Scholars know that the number of poems and prose composed by Li Qingzhao was enormous; however, only a miniscule fraction of her works still exist today. And among the total of 74 works attributed to her today, there are several for which the authorship is controversial, especially since the number of works attributed to Li Qingzhao has grown significantly over time. While Li Qingzhao was still alive, there were only 29 songs in her anthology. By the end of the Song dynasty that number grew to 35, and by 1550, there were 56 works attributed to her. Finally during the Ming dynasty, her anthology came to a halt at 74 individual works (Egan xxiii). Experts are skeptical of the authenticity of these works later added to her anthology, not only because they were lost for hundreds of years before resurfacing, but also because of the writing style differences between the works originally attributed to Li Qingzhao in the 12th century and some of the later attributed works. The question is: can we, as the modern reader, identify if Li Qingzhao is the actual author of these poems? Can computational techniques help determine authorship? How can we interpret the results of these computational techniques to help analyze her anthology? What are the limitations of using this technology on ancient Chinese texts? How can we optimize the algorithm to differentiate among 11th century Song dynasty poetry? These are some of the questions that this thesis aims to answer.

The first chapter provides a better understanding of the events in Li Qingzhao's life that shaped her place in history and her style of poetry. This chapter expands upon the topics in her anthology and how scholars relate her life events to her poetry, and the uniqueness of poetic

styles. This will help to understand the poetry in her anthology, the broader scene of Song poetry, and her contemporaries' writings. The section also examines the development of her anthology. More specifically, it researches how and when the poetry came to be included in her anthology. This is important for determining the authenticity of her poetry. And finally, this section provides a short analysis of a poem that is possibly a misattribution.

The second chapter investigates the use of machine learning and computational techniques of authorship identification and in particular, its applicability to pre-modern Chinese texts. Machine learning is a complicated topic, so this section aims to provide the reader with a better sense of what machine learning is, its limitations, and how it is used to determine authorship. By discussing current and past examples of how machine learning, natural language processing, and computational linguistics are used to help determine authorship, this section provides a breakdown of the specific processes involved in authorship identification, a critical component of the actual case study relating to Li Qingzhao. Previous examples provide a framework for analyzing texts. I develop and improve some of the algorithms for this project, as well as acknowledging some of the limitations of the algorithms; thus providing useful information on the effectiveness of this case study.

The third chapter discusses the actual implementation of the algorithm. This section discusses individual poems which were flagged by the algorithm as suspicious and presents estimations behind their authorship. More specifically, eleven poems are analyzed in context with their histories and the output of our computational test. This is where we make judgments about whether the poem is an imitation, a simple misattribution, or likely authentic. The poems are presented in a human-computer manner where the algorithm helps us judge each poem’s authenticity. This section also includes a more detailed analysis of these individual poems, based

on their historical background and combining human intuitiveness with computational robustness.

Finally, in the conclusion, I interpret these results together for a more complete understanding of Li Qingzhao and the works attributed to her. It discusses the results and how they compare to previous literature about the validity of her authorship. It highlights the significance of Li Qingzhao's authorship, the strengths and limitations of traditional and computational methods of authorship identification, and how these techniques might be useful for other types of similar authorship identification analysis in the future. Overall, this section concludes the case study and provides the reader with a greater understanding of authorship, Li Qingzhao, and computational abilities.

Can we use computational techniques to help identify the authorship in Li Qingzhao's anthology? Could publishers use Li Qingzhao's name to sell books written by others? Is it possible that other writers are trying to use her name to have their poems gain fame? What implications does our analysis of Li Qingzhao's anthology have on the field of poetry, computational linguistics, or the humanities as a whole? Will we be able to solve any mysteries? These are all very complex questions and some may not have an immediate answer, but a better understanding of Li Qingzhao (her anthology, her significance, and her authorship) and computational science will be attained through this case study. Perhaps these techniques can be used with other ancient texts with anonymous or multiple authors such as *Dream of the Red Chamber* 红楼梦 or *The Golden Lotus* 金瓶梅 to gain more insights on their authorship. These computational techniques may pick up clues about the past that humans are unable to perceive.

Chapter 1: Framing Li Qingzhao's Life and Poetry

> "Li Qingzhao is unquestionably the most
> celebrated women writer in Chinese history"
> (Egan ix).

This thesis contains an investigation into the authorship behind Li Qingzhao's 李清照 (1084–*ca.*1155) anthology of poetry utilizing computational methods. However, before diving into the theory behind the computer programming used for this analysis, we must better understand Li Qingzhao's background, her poetic style, and the context surrounding the addition of the later works in her anthology. Women writers and their lived experiences are an essential part of Chinese authorship. As Kang-I Sun Chang, says in her edited anthology *Women Writers of Traditional China*, "The lived situations of these women became paradigmatic for the Chinese understanding of authorship" (Chang 3). In this chapter, we explore Li Qingzhao's life, the social-political landscape of the Song dynasty, *ci* poetry, alterations to her anthologies over time, and an example of one of the controversial poems in her anthology. Her personal background and the context surrounding the growth of her anthology (the addition of poems to her famous anthology for hundreds of years following her death) is crucial in analyzing the poetry in her anthology, and it provides a preliminary non-computational framework for how to verify her authorship.

Li Qingzhao's Life

Li Qingzhao was born in the late Song dynasty in Jinan (in present day Shandong) into a high-class family of educated literati poets. Her father, Li Gefei 李格非 (*ca.* 1045–*ca.* 1105) was a student of Su Shi 苏轼 (1037–1101), one of the most accomplished poets and writers in

Chinese history. In Yao Dan's book *Chinese Literature*, Dan says, "Her father, Li Gefei, won recognition from Su Shi for the literary merit of his prose, and her mother Wang was also well educated. Li Qingzhao earned a name for her poems from childhood" (Dan 112). This recognition from Su Shi exemplifies the social status and merit of the family she was born into, as praise from Su Shi was a great honor. "Zhao Buzhi 趙補之(1053 – 1110) [one of Su Shi's friends] is said to have noticed the literary talent of Gefei's daughter, Li Qingzhao, at an early age, and remarked about it several times to other literary gentlemen of the time" (Egan ix). This would have been a great honor as well, clearly indicating the talent of young Li Qingzhao. Her skill was not only very developed at a young age, but also her political thought was mature and wise. She also wrote a few poems on the restoration of the Tang dynasty at the age of seventeen. Egan writes, "That a young girl could have produced such works was unheard of and attracted comment at the time, though no one says her poems are actually superior to the others [male poets of the time], which we would say today" (Egan x). Li Qingzhao's poetic aptitude was demonstrated and acknowledged at a young age, uncommon for a woman of this era, thus highlighting her place in history.

An arranged marriage started a new chapter in Li Qingzhao's life in 1101. At the age of eighteen, she was married to a national academy (Taixue 太学) student named Zhao Mingcheng 赵明诚 (1081-1129). He was a famous scholar and antiquarian who authored the *Catalogue of Inscriptions on Metal and Stone*[1], which records his lifetime collection of bronze and stone inscriptions with his own annotations. Zhao Mingcheng also came from a strong background, with his father working as a vice minister in the Song court. The politics of the time were

[1] *Catalogue of Inscriptions on Metal and Stone* 金石录 is an important catalog of Zhao Mingcheng and Li Qingzhao's collection of inscriptions, for which Li Qingzhao wrote the famous "Afterword"

turbulent, with different factions rising in power. Zhao Mingcheng's father was promoted to grand councilor, but this did not last. In 1107, Zhao Mingcheng's father was removed from office by a powerful minister. A few months later, Zhao Mingcheng's father died. This untimely death had lasting implications for Li Qingzhao and Zhao Mingcheng. It was customary for the children of recently deceased parents to observe a three-year mourning period in their family's ancestral home (in this case, Qingzhou in central Shandong). This mourning period interrupted Zhao Mingcheng's recent official career; however, it resulted in the couple living in Qingzhou for fourteen years and a long period of marital bonding. While living in Qingzhou, they collected books, rubbings, calligraphy, paintings, and ancient vessels— paying special attention to books and rubbings of inscriptions (Egan xi). Zhao Mingcheng and Li Qingzhao's collection of rubbings became the largest collection of rubbings ever compiled at that time, thus demonstrating the passion and determination behind this family's intellectual pursuit. Not only would they collect lots of art, inscriptions, and poetry, but they would also comment on the inscriptions— further adding to the value of their collection (Egan xi).

Of course, Zhao Mingcheng was impressive at the time, and famous in the following centuries, but Li Qingzhao had impressive merits of her own. In a passage written by Li Qingzhao in the "Afterword" of her husband's *Catalogue of Inscriptions on Metal and Stone*, she recounts a game they used to play that demonstrates her impeccable memory. She writes that they would often play a game where they would "point to a pile of books and, choosing a particular event, try to say in which book, which chapter, which page, and which line it was recorded" (Egan 75). Li Qingzhao then says that she would win more frequently than her husband, even though her husband spent years in the national academy studying and had a career dependent on his literary skills. In other words, even though Li Qingzhao was without an

academy education, she surpassed her husband in some aspects of literary achievement, although it was in the privacy of their own home. (Ronald Egan remarks that Li Qingzhao found it humorous that she uprooted the hierarchy of male-female literary skills.) These times of collecting and exploring literature in Qingzhou ended with the Jurchen invasion in 1125–26. Li Qingzhao and Zhao Mingcheng had to abandon a large part of their collection during their escape, but they still rented a "string of boats" to carry many of their possessions across the Huai River. Once they relocated to Jiankang (today's Nanjing), they learned that their collection of books in Qingzhou had been burned by the Jurchen. Their only remaining books in Jiankang became an obsession of theirs.

The following years would mark a transition from peaceful bliss surrounded by literature to a fast-paced and stressful few years. In 1128, Zhao Mingcheng was appointed the governor of Jiankang, a large and important city in the Song dynasty. Although prestigious, this appointment was also dangerous. In the year following his appointment, a plot to overthrow the government in the city was planned, causing Zhao Mingcheng to abandon his post with his own safety in mind. When this armed uprising failed, Zhao Mingcheng was disgraced and removed from his post. Following this incident, Li Qingzhao and Zhao Mingcheng sailed along the Gan River to continue their life in a safe location, away from conflict. Li Qingzhao must have been happy to return to a life with Zhao Mingcheng, but Zhao Mingcheng's fate changed again. Surprisingly, the Song emperor Gaozong 高宗 (r. 1127-1162) offered Zhao Mingcheng redemption by appointing him prefect of the city Huzhou. To demonstrate his gratitude, Zhao Mingcheng quickly left Li Qingzhao on the river and rode on horseback to give thanks to the emperor (the idea being that Li Qingzhao would follow him at a slower pace). To put this in perspective, he left her in an unknown region alone under the potential danger of being raided by the Jurchen

people. Li Qingzhao recounts this in her "Afterword" quoting Zhao Mingcheng, "Go with the crowd. If you must, discard the household belongings first, then our clothes, then the books and paintings, and then the ancient vessels. But the ritual vessels, be sure to take them with you wherever you go. Live or die with them. Don't forget!" (Egan 79). While abandoning his wife, he seemed to have more regard for their collection of ritual vessels than his wife in this instance.

Unfortunately for Li Qingzhao and Zhao Mingcheng, while returning to Jiankang he fell ill. Li Qingzhao heard of this news and rushed to see him, only getting a few days with him before he passed away at the age of forty-nine. This death was difficult for Li Qingzhao. Zhao Mingcheng left her with no support when he died, no method for income, or guarantee of safety. During the following months, Li Qingzhao travelled the south to avoid the Jurchen armies. These times were extremely difficult for Li Qingzhao. In a letter submitted to Hanlin academician Qi Chongli, it is explained that "all sorts of persons (Song military men, bandits, one of her temporary landlords, etc.) began to take advantage of her situation to pilfer or simply appropriate whatever portions of her extensive collection they could" (Egan 61). Li Qingzhao was very susceptible to danger. She had been married almost thirty years, never had a child, and her relatives were all dead, so there was nobody to take care of her, leaving the burden of survival all on her. Remarriage was relatively common at this time compared to the late imperial period and a husband would solve many of her problems. Because of this, three years after the death of her husband, she was remarried to Zhang Ruzhou 张汝舟, who was a low-ranking military officer. Although history does not know a lot about this man, it is known that Li Qingzhao sued him for misconduct in office— basically alleging that he was not qualified as a military officer. Apparently, "she was tricked into accepting Zhang's proposal by his sweet words and misrepresentation of his status and intentions" and "Zhang's real intent was, just like so many

other men she encountered after being widowed, to help himself to her wealth. When she would not give it to him, he began to beat her daily" (Egan xvi). She sued him because women could not directly file for divorce at this time, so she had to find other methods to separate herself from him. Due to the lawsuit, he was stripped of his position and sent into exile, resulting in their three-month long marriage to be annulled. Although she was freed from her abusive husband, she felt that she could not face members of high society again. She instead vowed to spend the rest of her life in solitude. It was during this time that she did a majority of her public writings— using her writing skills as a way to interact with high society and to restore her reputation. Clearly these times were difficult for Li Qingzhao, and little else is known about her during this time aside from her secluded life on rivers and lakes until she died. There are many unknowns surrounding the end of her life in regards to her social status and writings, but scholars debate whether she regained respectability or observed the life of seclusion in her last fifteen years of life.

As described above, Li Qingzhao lived a long life that was split into two periods: a time of bliss with her husband Zhao Mingcheng and a time of sadness after her husband died which overlaps with the disruption of Song dynasty history. Many of these feelings were captured either in her "Afterword" or in her poetry. Many experts use some of the events in her life to decipher complex metaphors within her poetry. This background is very helpful in our understanding of her poetry, her style, and her significance.

Poetic Form

Li Qingzhao is most famous for her *ci* poetry, a unique poetic form that is usually set to a tune. This sub-genre of poetry became more popular and was taken more seriously in the Song

dynasty. Yao Dan explains "*ci* , originally referring to the words of a song and called 'words to a tune' during the Five Dynasties, are lyrics that are supposed to be set to music" (Dan 102). At the early stage, these songs would be written by either men or women in a typically feminine voice since they would often be sung by women to the tune of music. The structure of *ci* poetry has a few forms, but is typically composed of lines of varying lengths. Dan explains, "Each *ci* has a fixed tune of its own. Mostly divided into the upper part and the lower part, the lines of *ci* are different in length. Therefore, *ci* is also called 'short and long lines'" (Dan 102). The themes within *ci* poetry vary, but they are "usually limited to writing about love affairs and women's sorrows. It is often flowery in language and sorrowful in artistic style" (Dan 102). During the Song dynasty, *ci* poetry became popularized. Ronald Egan explains, "In Li Qingzhao's time, the song lyric was still often written to be performed at parties and banquets at all levels of urban society, but it was increasingly also being honed by leading poets as a form of personal expression" due to the efforts of Su Shi (Egan xxii). This marked a change from traditional styles of poetry to a more vernacular style, where common people could understand more of the poems.

During this time, there was a debate on how *ci* poetry should be written. In the beginning of the twelfth century, scholars debated what exactly *ci* poetry is. Li Qingzhao had the famous position that the genre of *ci* poetry "is a different family" from other types of poetry (Chang 348). In Li Qingzhao's "Discourse on a Song Lyric", she expands upon this view, criticizing the *ci* poetry of notable authors. She explains in her "Afterword" that "*shi* poetry distinguishes between 'level' and 'oblique' tones, [however] the song lyric distinguishes five notes. It also distinguishes five tones, six musical modes, and the difference between 'clear' and 'turgid,' and 'light' and 'heavy' syllables"[2] (Egan 59). To Li Qingzhao, *ci* poetry does not share the same

[2] Translated by Ronald Egan in *The Works of Li Qingzhao*

basic values as all classical literature (Chang 349). She then names some other authors who she thinks have improved the genre, although imperfectly. She says,

> "Later, Yan Shuyuan (Yan Jidao 晏幾道, 1038–1110), He Fanghui (He Zhu 賀鑄, 1052–1125), Qin Shaoyou (Qin Guan 秦觀, 1049–c.1100), and Huang Luzhi (Huang Tingjian 黃庭堅, 1045–1105) appeared, and they were the first to truly understand the genre. But Yan's works suffer from lack of narrative exposition, and He's suffers from inadequate substance and classical style. Qin cares only about emotions and has too few literary references. His works are like a beautiful girl from a poor family. Although she may be gorgeous and radiant, she will never have the bearing of a lady from an affluent and high-ranking clan. As for Huang, although he prizes literary allusions, his works have many defects. They are like jade that has blemishes, reducing its value by half" (Egan 59).

Other poets and scholars of the time such as Wang Zhuo disagreed with her. They wrote their own critiques and embraced their personal styles. However, while analyzing Li Qingzhao's anthology, it is important to keep Li Qingzhao's ideal style and her critiques in mind. Li Qingzhao's emphasis on song lyrics being a separate class of poetry is exemplified in her pursuit to make her *ci* poetry more vernacular, and suitable for performances.

Li Qingzhao excelled at writing *ci* poetry. Her *ci* poems were shaped by her lived experiences. More specifically, her "*ci* poems were divided into two periods": before the Jurchen invasion forced her and Zhao Mingcheng to relocate, and after (Dan 112). Obviously, before the Jurchen invasion, Li Qingzhao's life was much more stable and she was surrounded by her husband and their massive amounts of collected antiques. During this time her literary talent was strong, but she was not yet influenced by the tragedies that followed the Jurchen invasion. After

the invasion and her husband's death, she had more tenure as a poet and more difficult life experiences to reflect on. Her *ci* poems "mainly covered the theme of love and life, expressing inner emotional feelings from the perspective of a woman's sensitivity. Marked by veiled and gentle, graceful and lively feminine beauty, her *ci* poems won her the name 'founder of *ci* poetry in a subtle and implicit style'" (Dan 115). Many other poets followed in her footsteps by writing in her style.

Li Qingzhao's Anthology and Authenticity Issues

A majority of Li Qingzhao's original works have been lost over time. The existing works of Li Qingzhao survived for almost a thousand years. As Idema says in her book *A Guide to Chinese Literature*, "The preserved body of her writings is regrettably very small. Apart from some fifty *ci* and a handful of *shi* , [her anthology] also comprises a moving autobiographical document in the form of a postscript written for her late husband's collection of epigraphical materials, in which she recalls their happy marriage" (Idema 158). This postscript refers to the "Afterword" that outlines much of her life, and these poems are in her most recent anthology. However, not all of these poems were always included in previous editions. This huge increase in songs might include fakes, and scholars, like Ronald Egan, think that many of these songs are problematic.

For example, "In Li Qingzhao's day song lyrics were regularly excluded from a person's literary collection, and they circulated, if at all, in a separate collection" (Egan 92). The first anthology to contain Li Qingzhao's works is *Jades for Rinsing the Mouth* 漱玉词 and this anthology is estimated to have been printed and circulated in the 1130s or 1140s. Unfortunately, this anthology was lost and current scholars do not even know how many pieces were in it.

Ronald Egan, in his English translation of Li Qingzhao's anthology, explains some of the issues in determining the authenticity of these writings. He says, "the two collections of her writings that once existed, a literary collection of prose and *shi* poetry and a separate collection of song lyrics, were both lost within a few centuries. That loss had enormous and dire consequences" (Egan xxiii). However, in 1146, during Li Qingzhao's lifetime, Zeng Zao 曾慥 (1091-1155) compiled the *Elegant Lyrics for Music Bureau Songs* 乐府雅词 containing twenty-three song lyrics attributed to Li Qingzhao (Egan 93). These twenty-three song lyrics are considered to be the most reliable attributions to Li Qingzhao. Many of these twenty-three song lyrics are also included in subsequent anthologies of *ci* poetry.

The *Garden of Plums* 梅苑, compiled by Huang Dayu 黄大于 in 1129 is one of the earlier anthologies containing songs attributed to Li Qingzhao. However, there are many suspicions surrounding this anthology. Not only were none of the six song lyrics attributed to Li Qingzhao included within the *Elegant Lyrics for Music Bureau Songs* 乐府雅词 or any other anthology until 1583, but one of the six lyrics was also an obvious misattribution. One of the *ci* poems to the tune of "Yuzhu xin" 玉烛新 is attributed to Li Qingzhao in the *Garden of Plums*, but the same piece was also included in Zhou Bangyan's song lyric collection, predating *Garden of the Plums*. Ronald Egan in his book *The Burden of Female Talent* outlines the histories behind each of the additions to Li Qingzhao's cumulative attributions as well as their problems. He also includes a helpful chart displaying the attributions of each individual song lyric over time, emphasizing which song lyrics show up in multiple anthologies[3].

Later, in the Ming dynasty her anthologies became even more popular, and even more problematic. "In the Ming period alone, when compiling and printing Song dynasty song lyrics became a fad among publishers, there is a dismaying hash of works, many brought out in

[3] Page 96–97.

multiple recensions with some identical titles having different contents, and some different titles having the same contents" (Egan xxiii). Although there are a total of 74 poems attributed to her, many of them were added to her anthology over time. While she was still alive, there were only 29 poems attributed to her. By the end of the Southern Song dynasty, there were 35. By 1550, there were 56 poems. And by the end of the Ming dynasty, there were 74 poems attributed to her. Many of the additions occurred during the Ming dynasty and the potential for publishers to intentionally misattribute works as that of Li Qingzhao was very real. However, "on the other hand, we can readily think of reasons Ming and Qing editors might have *imagined* that previously unknown or unattributed pieces were written by Li Qingzhao (as her fame was growing)" (Egan xxiv). This scenario would benefit the computer program, because then people would not painstakingly imitate Li Qingzhao's works, and instead the poems would be less likely than careful imitations to pass as hers— effectively making it easier for the computer to spot a difference.

Authentic Song Lyric and Potential Imitation

This section includes two *ci* poems written to the tune of "Immortal by the River". The first poem (no. 3.19 in Ronald Egan's anthology of Li Qingzhao) is a poem that is considered most credibly written by Li Qingzhao. It was included in the most reliable anthology, *Elegant Lyrics for Music Bureau Songs* and it has since occurred in fifteen other anthologies prior to 1900. Attached to this song is a note by Li Qingzhao, which says "Master Ouyang (Ouyang Xiu) wrote a song lyric to the tune 'Butterfly Loves Flowers' with the line 'Deep, the deep courtyard, how deep is it?' which I'm most fond of [4]. I have borrowed his line to write several 'Deep, the

[4] Some scholars argue that Ouyang Xiu actually was not the author of this line. Instead, it was likely a product of Southern Tang poet Feng Yansi 冯延巳 (903-1060) (Egan 125).

deep courtyard' songs. In fact, the tune he used is the one formerly known as 'Immortal by the River"' (Li Qingzhao). This note not only mentions that there are multiple Li Qingzhao authored poems to the tune "Immortal by the River", but it also provides the first line of the poem, "Deep, the deep courtyard, how deep is it?". Interestingly enough, there is another "Immortal by the River" song lyric included in a different anthology of hers. This song (no. 3.57 in Ronald Egan's Anthology on Li Qingzhao), was included in the suspicious *Garden of Plums* which existed during Li Qingzhao's lifetime. However, in this anthology, it was not attributed to Li Qingzhao. It took another four centuries for this poem to be attributed to Li Qingzhao in the *Huacao cuibian* 花草粹编. The *Huacao cuibian* was published in 1583 and the Li Qingzhao attribution for this song lyric is potentially implicit (assumed to be Li Qingzhao) (Egan 96). It is skeptical that an unattributed poem was for around four hundred years before first being attributed to Li Qingzhao in her most comprehensive anthology of the sixteenth century. Did the publishers include it in the *Huacao Cuibian* because they wanted more attention for their anthology?

According to the note left attached to Li Qingzhao's "Immortal by the River" included in *Elegant Lyrics for Music Bureau Songs,* Li Qingzhao was fond of Ouyang Xiu's "Magpie Steps on the Branch" and used it's first line as inspiration for multiple of her poems. The piece is a potential misattribution and may have been written by Southern Tang poet Feng Yansi 冯延巳 (903-1060) and it includes the same first line "Deep, the deep courtyard, how deep is it?" (Egan 125). Ouyang Xiu's poem is written to the tune "Magpie Steps on the Branch" and is included below.

鹊踏枝

庭院深深深几许，

杨柳堆烟，

帘幕无重数。

玉勒雕鞍游冶处，

楼高不见章台路。

雨横风狂三月暮，

门掩黄昏，

无计留春住。

泪眼问花花不语，

乱红飞过秋千去。

To the tune “Magpie Steps on the Branch”

Deep, the deep courtyard, how deep is it?

Willows pile up the mist,

Countless layers of blinds and curtains

Jade bridle and carved saddle are in the pleasure quarters

Zhangtai road cannot be seen from the high tower

Wild winds drive slanting rain, sunset in the third month

The gates shuts in the dusk

But there’s no way to detain spring

Tear-filled eyes ask the flowers, the flowers do not speak

A whirl of red petals flies past the garden swing

(“庭院深深”)

Now that the inspiration for Li Qingzhao’s “Immortal by the River” poems have been taken into account, the two at-issue poems can be read. The Chinese and English-translated versions of these two poems are included on the next page, along with the note.

First Poem (*Elegant Lyrics*)

臨江仙

庭院深深深幾許

雲窗霧閣常扃。

柳梢梅萼漸分明。

春歸秣陵樹

人客建康城。

感月吟風多少事

如今老去無成。

誰憐憔悴更彫零。

試燈無意思

踏雪沒心情。

Note:

歐陽公作《蝶戀花》有「庭院深深深幾許」之句，予酷愛之，用其語作「庭院深深」數闋。其聲蓋即舊《臨江仙》也。

Second Poem (Garden of Plums & Huacao cuibian)

臨江仙

庭院深深深幾許

雲窗霧春遲。

為誰憔悴損芳姿。

夜來清夢好

應是發南枝。

玉瘦檀輕無限恨

南樓羌管休吹。

濃香吹盡有誰知。

暖風遲日也

別到杏花肥。

First Poem (*Elegant Lyrics*)

Deep, the deep courtyard, how deep is it?

Cloudy windows and misty halls are forever locked.

Willow tips and plum buds can gradually be seen.

Spring returns to the trees of Moling, this person is a sojourner at Jiankang city Moved by the moon, chanting in the wind, so much has happened!

Today I'm old and have accomplished nothing.

Haggard and declining, yet who shows concern?

Lighting the lanterns holds no interest for me, and I've no enthusiasm for treading on the snow (Egan 125).

Note:

Master Ouyang (Ouyang Xiu) wrote a song lyric to the tune "Butterfly Loves Flowers" with the line "Deep, the deep courtyard, how deep is it?" which I'm most fond of I have borrowed his line to write several "Deep, the deep courtyard" songs. In fact, the tune he used is the one formerly known as "Immortal by the River."

Second Poem (Garden of Plums & Huacao cuibian)

Deep the deep courtyard, how deep is it?

Cloudy windows and misty halls, late in spring. For whom are you so weakened, your fragrant beauty diminished?

Last night in my lovely dream you were fine, I thought you'd be filling the southern branches.

The jade is grown frail, the sandalwood hue faded, how sad! Don't let the Tibetan flute play its melody in the southern loft. When your fragrance is blown away who will know? The wind is warm, the days of sunshine long, and the apricot blossoms plump (Egan 184).

Trying to determine which poem is most similar to the other poems attributed to Li Qingzhao is quite a difficult task for a reader. The goal of this research is to distinguish authorship in Li Qingzhao's anthology. The two poems above are a unique example of a later attribution. By looking at the context described in the note and the writing in poems, it is possible that these are both from Li Qingzhao, but also possible that the second poem is a misattribution. Further analysis provides a deeper understanding for the obstacles of authorship identification.

These two poems highlight the uncertainty of authorship which is representative of many of Li Qingzhao's works. There are two poems written to the same exact tune and both inspired by Ouyang Xiu's poem, but also have some potentially illuminating stylistic differences. For example, the first poem is more vernacular and reads more like other Li Qingzhao works. More specifically, the first poem uses the terms 没心情 (no enthusiasm), 无意思 (no interest), 多少事 (so much has happened), and 谁 (who). These are more vernacular words that are known to be in Li Qingzhao's vocabulary based on reading other reputable works written by her. The second poem also has a few words that are more vernacular, but they are less frequent and also not used as smoothly. For example, the second poem has 谁 (who) and 梦好(lovely dream) which are more vernacular, but most of the terms in the second poem are more similar to more traditional forms of poetry.

The repeated characters in the first line are typical of Li Qingzhao's style, and these are common to both poems as indicated in the note. It is also known that Li Qingzhao travelled to Moling and Jiankang, so the first poem has the context of her life correct. She also mentions that she is old and has accomplished nothing, which seems to be a common theme in her later poetry. The usage of allusion within the two pieces are also important for analyzing the poetry. The

second poem also alludes to being old and accomplishing nothing with the line “When your fragrance is blown away who will know?”. However, this idea is presented in the second poem in a much more literary style, whereas in the first poem, Li Qingzhao expresses it in a more straightforward way and colloquial style.

These differences in style, when combined with historical knowledge about when each poem first surfaced tells an interesting story. It is clear that this poem is a reputable Li Qingzhao piece because it’s included in *Elegant Lyrics for Music Bureau Songs* and it’s style matches Li Qingzhao’s. The other poem was not included in the *Elegant Lyrics* anthology, and was initially included in another anthology *The Garden of Plums* which was printed in 1129 while Li Qingzhao was still alive. However, the poem was not attributed to Li Qingzhao in this anthology, and the first time this poem was attributed to Li Qingzhao was several hundred years later in a Ming dynasty anthology.

There are some other similarities and differences between these poems, and her other poems as a whole. Perhaps the note and other “Deep, the Deep Courtyard” songs were included within *Jades for Rinsing the Mouth*, making it very easy for people with access to this lost anthology to create convincing imitations. It is also possible that this was an actual work written by Li Qingzhao, just in her earlier days. When compared with Ouyang Xiu’s poem which inspired Li Qingzhao, the second poem shows more similarity than the first poem. The second poem and Ouyang Xiu’s poem both utilized the late spring scenery to suggest an under-achieving theme and decline. Poem one expresses similar sentiments much more candidly and less figuratively. Because of this, the second poem could be an earlier practice, in which Li Qingzhao was imitating Feng more carefully, and later in the first poem she developed her own style. Trying to better understand the authorship behind these pieces with only historical information,

style, and small notes is difficult. Perhaps some computational methods will help shed some light on the authenticity of these poems.

Chapter Two: Computational Methods

Motivation

Authorship verification of Li Qingzhao's anthology is quite a difficult goal. There are many components that go into determining the authenticity of poems within Li Qingzhao's anthology, ranging from historical information such as when the poems were first published to the more poetic aspects such as diction, style, tone, and rhythm. The historical analysis has been comprehensively studied by Ronald Egan and he proposes several different levels of credibility: most credible (1-23), more credible (29-35), less credible (24-28, 37, 38-45) and not credible (46-66) being problematic (Shields xxv). Egan says, "The criteria I am using to distinguish degrees of credibility are based entirely on the date and reliability of the earliest attribution as well as subsequent confirmation of the same or the lack of it" (Egan 104). He explains that he does not include any stylistic considerations because an imitator could also pick up on her style and write in her style, thus making it very difficult to distinguish authenticity from imitation. Although it is true that other master poets could have copied her style really well and it may be difficult for present day scholars to differentiate the real poems from imitations, there is research into how computational methods can help with determining authenticity. According to Patrick Juola's book *Authorship Identification*, "recent developments of improved statistical techniques in conjunction with the wider availability of computer-accessible corpora have made the automatic and objective inference of authorship a practical option" (Juola 2). An example of authorship attribution is Mosteller and Wallace's analysis of the *Federalist* papers, a famous set of newspaper essays published by an anonymous person. Monsteller and Wallace were able to analyze the diction of the words used within the papers to determine that John Jay wrote five,

James Madison wrote fourteen, and Alexander Hamilton wrote fifty-one (Juola 10). Although the circumstances behind this analysis and the case of Li Qingzhao are different, there is precedent for the use of computation within authorship identification.

Since there is a small number of poems within our corpus, and there are multiple potential authors, there are certain limitations for which computational methods can be used. Machine learning, for example, is a growing field that carries lots of excitement and potential. The computer algorithm studies a training set of texts known to be from certain authors and then the computer estimates authorship. There are accessible methods for doing this in popular programming languages such as Python or Mathematica. Wolfram Alpha, the creators of Mathematica created a model that only requires around ten lines of code to estimate the authorship behind *Macbeth*, *An Ideal Husband*, and *The Man Who Laughs* accurately as Shakespeare, Wilde, and Hugo ("Determining the Author of a Text"). There are entire libraries in Python dedicated to solve classification problems like authorship attribution. Although these methods are interesting and can be very accurate, Li Qingzhao's anthology is not so simple. In the article *Authorship Attribution in the Wild* by Moshe Koppel et al. the complexities of real world authorship attribution is explored. He explains how many of the studies done in authorship identification is when we are trying to match an anonymous document to a "small set of candidate authors" and goes on to name this the "vanilla" version (Koppel 84). Unfortunately for us, the case of Li Qingzhao is far from the "vanilla" version because there could be hundreds of candidate authors who may have written a poem included in her anthology, the author might be unknown, and the amount of texts from these authors is very limited. So as we can see, the authorship attribution of Li Qingzhao's anthology is fundamentally different from the models

created by Wolfram Alpha or many machine learning models within other programming languages.

Koppel says, "Broadly speaking, methods for automated authorship attribution can be divided into two main paradigms. In the similarity-based paradigm, some metric is used to measure the distance between two documents and an anonymous document is attributed to that author to whose known writing (considered collectively as a single document) it is most similar" (Koppel 84). Machine learning techniques can be very accurate at determining the authorship behind a text when there is lots of training data. However, since Li Qingzhao's anthology is limited in size and it is not considered a "vanilla" version of authorship attribution, a similarity-based attribution approach is more appropriate. Koppel says, "In the case that there are many authors, Koppel et al. (2006) and Luyckx and Daelemans (2008) have asserted that similarity-based methods are more appropriate than machine learning methods" (Koppel 84). Similarity-based methods measure how similar two documents are by using different mathematical approaches such as Jaccard similarity and cosine similarity. Although both Jaccard similarity and cosine similarity are able to measure the distance between two texts, Jaccard similarity does not factor frequency into the equation (Gupta). This means that if a sentence contains duplicate words, then cosine similarity is a better measure of similarity for the two texts.

Theory Behind Cosine Similarity

Cosine similarity allows us to have a numerical metric for the similarity of the words used between certain texts. The theory behind implementing this method on Li Qingzhao's anthology is that the poems that are least similar or most similar to other poems within her anthology will be flagged as suspicious. The poems that show less similarity are either very

unique poems written by Li Qingzhao or less convincing imitation poems. On the other hand, poems which show a lot of similarity are either authentic Li Qingzhao poems, or convincing imitation poems.

This section talks about the theory behind implementing cosine similarity on Li Qingzhao's poems. The actual process of finding the cosine similarity can be broken down into different sections. First, we have the text preprocessing where the poems are digitized, punctuation is removed, and the words are tokenized into chunks. The second step is to create a list of all the words used in the poems, transform each poem into vectors (term frequency bag of words or term frequency inverse document frequency bag of words), and then compute the cosine similarity. The third section is to input these results into a matrix, display them visually, and compute summary statistics which help us differentiate each poem's similarity.

Text preprocessing is the process of cleaning all the data within the poems so that we can input them into our model. The poems were taken from Ronald Egan's *The Works of Li Qingzhao* and compiled into an Excel file with each line containing the poem number, the name of the poem, and the text. The Excel file is then read into Python where more preprocessing occurs. Typically, when cleaning the text in English there are several operations that we need to do so that we can create a vector out of the words. For example, we will look at three sentences and the preprocessed text. The table below shows the sentences and the tokenized & preprocessed text.

Sentence number	Before Processing	Tokenized & Preprocessed
Sentence 1	I want to eat lunch sooner	[want, eat, lunch, soon]
Sentence 2	Let's get Lunch soon?	[lunch, soon]

Sentence 3	How much is Lunch?	[much, lunch]

Table 1: Example Sentences and Tokenized Versions

We lower all of the cases so that they are uniform across all sentences, remove all punctuation and get rid of suffixes. The algorithm works by matching exact words across sentences. For instance, notice how the word “sooner” gets shortened to the word “soon”. This process is called lemmatization and is less important for Chinese. We also remove stop words, or the words within the sentence that carry little meaning, leaving just the words that carry more meaning within the sentences. Stop words can vary depending on the analysis. The tokenization is the process of splitting up the words into different elements within a list. In English, this is usually simple since most words are separated by a space.

After tokenizing the sentences, we are able to create a list of all the important words across the three sentences giving us the list: (want, eat, lunch, soon, much). Using the words in this list, we vectorize the sentences by placing the frequency of the words common to the sentence and the bag of words list within a vector. This process is called the term frequency bag of words because we are only counting the number of occurrences of each word. Since there are only zero or one occurrences in the sentences below, we get the three vectors below.

Words:	want	eat	lunch	soon	much
Sentence 1:	1	1	1	1	0
Sentence 2:	0	0	1	1	0
Sentence 3:	0	0	1	0	1

Table 2: Example Sentence Vectors for Bag of Words

Writing them in vector form, sentence vector one is [1,1,1,1,0], sentence vector two is [0,0,1,1,0], and sentence vector three is [0,0,1,0,1].

We can also compute the TF-IDF (term frequency-inverse document frequency) bag of words vectors as well. The TF-IDF method requires us to normalize the frequency by dividing by the total number of like terms across all sentences. We can divide each observation by the number of terms in the sentence. For example, all of the frequencies in the first sentence are divided by 4, whereas the frequencies in sentence two and sentence three are divided by 2. When we do this, we get the normalized table below.

Words:	want	eat	lunch	soon	much
Sentence 1:	1/4	1/4	1/4	1/4	0
Sentence 2:	0	0	1/2	1/2	0
Sentence 3:	0	0	1/2	0	1/2

Table 3: Example Term Frequency Vectors for TF-IDF Method

Next, we need to multiply each of these elements by the inverse document frequency (IDF). This is found using the equation (Huilgol):

$$idf_t = \log \frac{number\ of\ documents}{number\ of\ documents\ with\ term\ 't'}\ .$$

In our example, we can calculate IDF('want') = $\log \frac{3}{1} = .477$ since there are three documents and only one occurrence of the word 'want'. Similarly, we find that the other inverse document frequencies are as follows:

	want	eat	lunch	soon	much
IDF	$\log\frac{3}{1} = .47$	$\log\frac{3}{1} = .47$	$\log\frac{3}{3} = 0$	$\log\frac{3}{2} = .23$	$\log\frac{3}{1} = .47$

Table 4: Example Inverse Document Frequency Vectors for TF-IDF Method

Now we have an IDF vector which we can multiply by our normalized frequency vectors by the inverse document frequency vectors to get the resultant TF-IDF vectors. We get the following table:

Words:	want	eat	lunch	soon	much
Sentence 1:	$1/4 \cdot .477$	$1/4 \cdot .477$	$1/4 \cdot 0$	$1/4 \cdot .238$	$0 \cdot .447$
Sentence 2:	$0 \cdot .477$	$0 \cdot .477$	$1/2 \cdot 0$	$1/2 \cdot .238$	$0 \cdot .447$
Sentence 3:	$0 \cdot .477$	$0 \cdot .477$	$1/2 \cdot 0$	$0 \cdot .238$	$1/2 \cdot .447$

Table 5: Example TF-IDF Vectors for TF-IDF Method

Writing them in vector form, the TF-IDF vector for sentence one is $[0.119,\ 0.119,\ 0,\ 0.059,\ 0]$, the TF-IDF vector for sentence two is $[0,\ 0,\ 0,\ 0.119\ , 0]$, and the TF-IDF vector for sentence three is $[0,\ 0,\ 0,\ 0,\ 0.2385]$. When comparing the bag of words approach and the TF-IDF approach, we see that words that are common to all of the sentences are weighed as zero and that the words that occur in most of the sentences are weighed less heavily. The inverse frequency approach emphasizes uncommon words when there is lots of data. This is important within our algorithm because it weighs the less common words more heavily than the words that occur in almost every poem.

Now, we have our bag of word vectors and our TF-IDF vectors for each sentence, so we can compare them by using the cosine similarity function. Cosine similarity is a way to measure the similarity of two vectors by using an inner product. It utilizes the formula for the Euclidean dot product:

$$A \cdot B = ||A||\,||B||\,cos\,\theta$$

where A and B are non-zero vectors and θis the angle between the two vectors. By rewriting the formula to solve for $cos(\theta)$, a measure for the similarity for the two vectors is found. This cosine similarity formula is given below:

$$similarity = cos(\theta) = \frac{A \cdot B}{||A||\,||B||} = \sum_{j=1}^{n} \frac{A_j B_j}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}.$$

Typically this formula would be used inside the linear algebra classroom, but it has also been a tool within natural language processing. An article written by Varun Chaudhary, an industry expert on natural language processing, explains that "cosine similarity is one of the most widely used and powerful similarity measures in Data Science. It is used in multiple applications such as finding similar documents in natural language processing, information retrieval, finding similar sequences in DNA bioinformatics, detecting plagiarism and many more" (Chaudhary). In our case, we will be using cosine similarity to compare the vectors created using bag of words and TF-IDF natural language processing techniques.

The process for calculating the cosine similarity of the vectors is the exact same regardless of whether we use term frequency bag of words or TF-IDF bag of words. To show a simple example, we can calculate the cosine similarity for the term frequency bag of words vectors for sentence 1, sentence 2, and sentence three. The vector for sentence one is

[1, 1, 1, 1, 0], the vector for sentence two is [0, 0, 1, 1, 0], and the vector for sentence three is [0,0,1,0,1]. The cosine similarity between the three vectors is calculated below.

$$similarity\ one\ and\ two\ = \frac{1\cdot0 + 1\cdot0 + 1\cdot1 + 1\cdot1 + 0\cdot0}{\sqrt{1^2+1^2+1^2+1^2+0^2}\times\sqrt{0^2+0^2+1^2+1^2+0^2}} = \frac{2}{2\sqrt{2}} = .707.$$

$$similarity\ one\ and\ three\ = \frac{1\cdot0 + 1\cdot0 + 1\cdot1 + 1\cdot0 + 0\cdot1}{\sqrt{1^2+1^2+1^2+1^2+0^2}\times\sqrt{0^2+0^2+1^2+0^2+1^2}} = \frac{1}{2\sqrt{2}} = .354.$$

$$similarity\ two\ and\ three\ = \frac{0\cdot0 + 0\cdot0 + 1\cdot1 + 1\cdot0 + 0\cdot1}{\sqrt{0^2+0^2+1^2+1^2+0^2}\times\sqrt{0^2+0^2+1^2+0^2+1^2}} = \frac{1}{2} = .5.$$

According to our process, the first and second example sentences are more similar to each other than the first and third sentences. This makes sense because the first two sentences are talking about having lunch soon, whereas the third sentence is talking about the cost of lunch. The reason that the similarity between two and three is higher than the similarity between one and three is simple. Sentence one had more words within its bag, containing "want, eat, lunch, soon", so each similarity is weighed less. Whereas sentence two only has two words in its bag, so the common words in sentence two and three are weighed heavier.

We can also create a matrix of cosine similarities between all the sentences. For example, we see that the table below compares all of the sentences with each other so that we can easily see how each sentence compares with the others.

	Sentence 1	Sentence 2	Sentence 3
Sentence 1		Similarity of 1 and 2	Similarity of 1 and 3
Sentence 2	Similarity of 1 and 2		Similarity of 2 and 3
Sentence 3	Similarity of 1 and 3	Similarity of 2 and 3	

Table 6: Example Cosine Similarity Matrix

The diagonal is intentionally left blank because it would represent the similarity between each sentence and itself, so it always has a value of one. Filling in the bag of words cosine similarity values for the three example sentences, we get the matrix below.

	Sentence 1	Sentence 2	Sentence 3
Sentence 1	1	.707	.354
Sentence 2	.707	1	.5
Sentence 3	.354	.5	1

Table 7: Example Cosine Similarity Matrix with Values

We can also represent this matrix visually by using a correlation heatmap matrix which uses darker shades of colors to show where the correlation is higher. For example, the number .707 would be darker than .5 which would be darker than .354. Thus showing any patterns in the data.

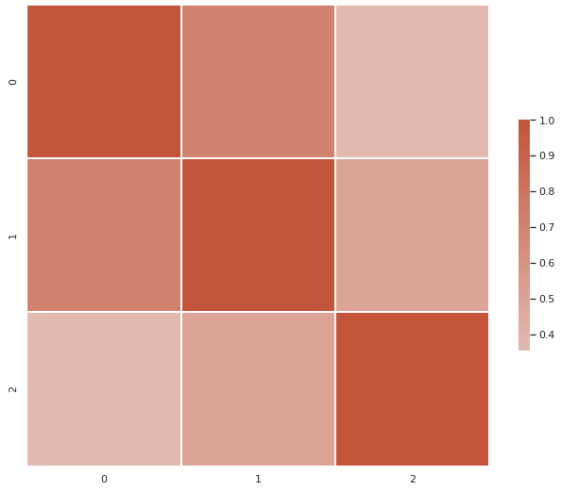

Figure 1: Example of Similarity Heatmap

The first column (indexed as zero) represents the cosine similarities for the first sentence, the second column represents the second poem, and the third column represents the similarities for the third sentence. If we were to sum the non diagonal tiles in each column, we would find the cumulative similarity for each sentence. For example, cumulative similarity for sentence one is $.707 + .354 = 1.061$.Sentence two is $.707 + .5 = 1.207$ and sentence three's total correlation is $.354 + .5 = 0.854$. According to this metric, we would say that sentence two is the most similar sentence within the three, with sentence three being the least similar.

Modifications for Chinese Texts

Working with three short sentences demonstrates the process behind the theory of cosine similarity with a bag of words. We can implement this same process for Li Qingzhao's poems, where each poem has an associated bag of words that we can compare. Then, the result of the heatmap and total correlations would show which poems relate to each other, and by how much. In essence, it shows the similarity in the diction used in Li Qingzhao's anthology. Although there are many other aspects of poetry that can be used to determine authorship, a comprehensive analysis of the diction behind her poetry is a good place to start. It can help point to lexically unique poems, and more standard poems within her anthology. We may find that some of the more unique poems also do not exhibit much Li Qingzhao style, or that historically speaking, they are unlikely to be hers. In the experiment we will rank her poems on similarity using multiple metrics and methods. Since we are looking at ancient Chinese poetry, there are many differences in the methods outlined in the example above, and the methods employed in our experiment. First, let's look at a stanza from one of Li Qingzhao's poems. Looking at poem

number 3.16 (鷓鴣天), a poem included in Li Qingzhao's more reliable anthology. We see that the first stanza is as follows (Egan 118):

寒日蕭蕭上鎖窗。

梧桐應恨夜來霜。

The first step in our process is tokenization, where we separate all of the words into their respective groups. Within Python, there is a module for natural language processing in Chinese called Jieba. Jieba's focus is on contemporary Chinese which differs heavily from the language used in premodern *ci* poetry. Since Jieba does not have the capability to properly tokenize premodern Chinese, we end up with a tokenized stanza of

[寒日, 蕭蕭上, 鎖窗, 梧桐, 應恨, 夜來]

where the commas separate each tokenized word. Immediately, we notice that the last character frost (霜) is missing from the tokenized section. However, this word does contain meaning and is important. The translation of this stanza is

"The cold sun is bleak, climbing the lattice window.

The paulownia tree must resent last night's frost." (Egan 119).

Another issue in the Jieba text preprocessing package is that characters such as 上 (meaning "to rise", "on", or many other things depending on the context), can be used in conjunction with another word to modify it, or can be used on its own as a verb. In this case, Jieba thinks 上 is being used in conjunction with 蕭蕭 (meaning "slightly cold and desolate") so it tokenizes the phrase as 蕭蕭上. However, it makes more sense that the word 上 is being used to describe the action of the sun (the sun is rising). Since 上 is actually being used as a verb on its own, describing the action of the sun, it should be segmented as "蕭蕭, 上" as two different word chunks. Because of these issues with Jieba, we must be wary of the results of the tokenized

text. If we use Jieba's text processing package then we are bound to lose some information, and have some mistakes splitting up phases. This influences the bag of words because it misses some potentially important words, or groups them incorrectly and matches them wrong. For example, if one poem contains the phrase "蕭蕭" and another has the phrase "蕭蕭上", the cosine similarity function will count these two phrases as different, when in reality they should be similar.

Since Jieba, the main Chinese text processing module in Python has some issues, it is important to come up with alternative methods for text preprocessing (the method of breaking up the sentence into chunks before the bag of words is created). In this thesis, two alternative methods are explored. The first method, tokenizes the words by characters. In Chinese, each character has an attached meaning, and when we group these characters together, we get unique words. We can break these words down into their smallest components and see what types of characters are most frequently used. This way, each character has its own weight attached to it. For example, we would tokenize the stanza as

[寒, 日, 蕭, 蕭, 上, 鎖, 窗, 梧, 桐, 應, 恨, 夜, 來, 霜].

This method gives us a lot more similarities than the Jieba method because almost all poems are going to contain at least a few similar words to other poems. Although this method does not have the issues associated with the Jieba method, it has its own issues. First, almost all of the poems have several of these characters within it, making the distribution of similarities more uniform. This makes it more difficult to find poems that are either outliers or on either side of the average cosine similarity value. To help mitigate this, we can use the TF-IDF method instead of bag of words because TF-IDF weighs more common words less heavily than words with a lower frequency.

These first two methods are less subjective because they rely on standardized tokenization techniques. This broad approach may be beneficial for many applications; however, their unspecialized methods have additional issues. For example, in the Jieba method, we showed that sometimes it drops important words, or groups characters together in a way that is different from the way it should be read poetically. The second method, tokenizing by character also has some issues, where it does not pick up words that are meant to be next to each other. These chunks of multiple characters make up words and also compose words that are a fixed collocation (固定搭配). For example, the characters 寒 (cold) and 日 (sun) have their own meanings, but when we put them together we get 寒日 which describes cold weather or winter's sunlight and holds special meaning within Chinese poetic tradition.

Words like "cold sun" (寒日) have been used by other great poets like Tao Qian 陶潜 (365 - 427). For instance, in his poem *In Reply to Aid Pang* 答庞参军, Tao Qian uses "cold sun" to describe a desolate, lightless, and miserable scene. Tao Qian was a very influential poet, whom Li Qingzhao most likely would have read previously. As discussed in Chapter One, she believes ideal *ci* poetry uses references to other poems in a balanced fashion. Whether or not she copied Tao Qian directly, such textual references would enhance the emotional density and literary quality of the poem because a learned reader would reflect on earlier canons to make sense of the current topic.

To prevent these issues, it makes the most sense to actually Tokenize these poems by hand. It takes an expert to be able to split these poems up into the important chunks that carry enough meaning. This third method is more subjective, but most accurately tokenizes the poems in a way more similar to how they are supposed to be understood. Dr. Chao Ling individually tokenizes each poem. This way, words such as 寒日 which hold poetic meaning are preserved,

and phrases such as 蕭蕭上 are tokenized correctly. For example, he tokenized the first stanza in Li Qingzhao's poem number 3.16 (鷓鴣天) as: (寒日, 蕭蕭, 上, 鎖窗, 梧桐, 應, 恨, 夜來, 霜). The appendix contains an excel file that has all of Ling's tokenized poems in Li Qingzhao's anthology. Others may feel that some of the poems should be tokenized in a different way, but for the remainder of this thesis, we will assume that Dr. Ling's tokenization is the closest to how traditional readers would have read Li Qingzhao poetry.

Experiment Using Dr. Chao Ling's Tokenization

The motivation behind this case study is to investigate the diction within the poetry included in Li Qingzhao's anthology and to see if we can make any claims on the authenticity of the authorship behind the poetry using cosine similarity with bag of words. We include the 66[5] song lyrics that are included in Ronald Egan's anthology of Li Qingzhao in our case study. We run the cosine similarity algorithm using the different methods discussed in the previous section (term frequency bag of words and TF-IDF using the different tokenization methods). In this section we compute the term frequency bag of words cosine similarity for Ling's tokenization. The results of each test, along with some interesting findings and theories are discussed.

The words that show up the most frequently using Ling's tokenized poems are

'誰': 6, '醉': 6, '盡': 6, '似': 6, '未': 7, '上': 7, '黃昏': 7, '愁': 7, '恨': 7, '到': 8, '又': 8, '不': 8,

'無': 8, '莫': 8, '憔悴': 8, '深': 9, '來': 10, '好': 10, '更': 11, '春': 11, '風': 12, '人': 20

where the numbers next to the words represent the frequency of the word throughout all of the poems. Many of these words are descriptive words that are common to many Chinese poems and other words are used less than six times throughout all the poems. Since these words are not

[5] There are 74 total poems, but 8 of them are universally considered misattributions.

common to a majority of the poems, it is more appropriate to use the term frequency bag of words model instead of the TF-IDF model. Using this word frequency list to compute the term frequency bag of words, we obtain the following correlation matrix.

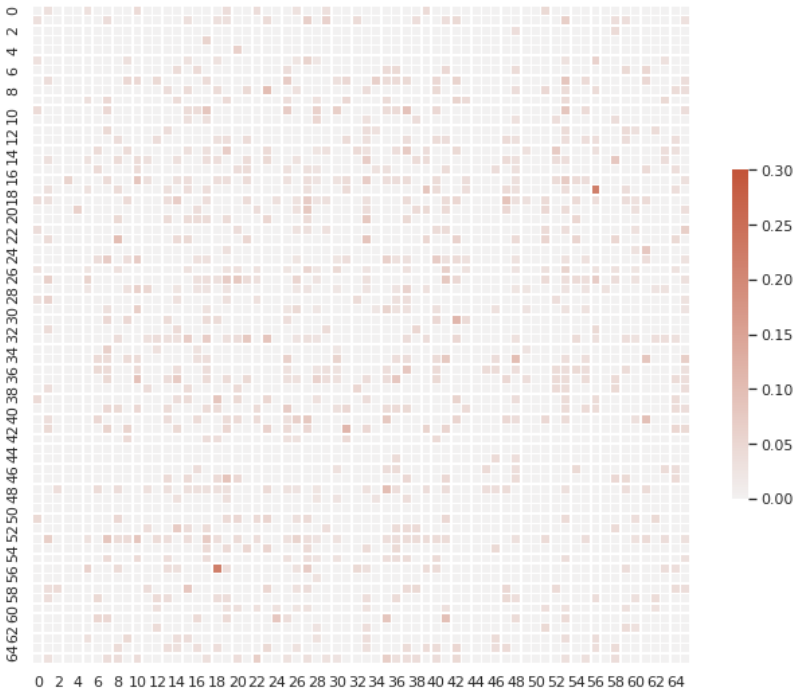

Figure 2: Heatmap of Cosine Similarity with Dr. Chao Ling's Tokenization

Notice how there is low correlation between all of the poems. With these darkest squares representing the highest correlation between individual poems. The strongest similarity in the graph shown above is between poem number 3.19 and poem number 3.57, the two poems discussed in Chapter One which are based off Ouyang Xiu. This makes sense because the first line of each poem is exactly the same. Aside from these two poems, we can look at the poems with the lowest maximum similarity, meaning that they are not very similar to any other poem in the anthology. The poems with the smallest maximum similarity are as follows: poems 3.4, 3.45, 3.5, 3.58, 3.64, and 3.3. There are two other metrics that are used to look at the similarity across

poems. One metric is the non zero similarity count. This is the number of poems that share similar words with the specific poem. A list of the poem number, name, maximum similarity, and non-zero similarity count is provided below for the lowest six similarity scores.

Poem Number	Name of Song	Maximum Similarity	Non-zero Sim Count
3.4	如夢令	0.048	7
3.45	春光好	0.049	17
3.5	如夢令	0.052	21
3.58	山花子	0.059	13
3.64	品令	0.059	28
3.3	漁家傲	0.060	21

Table 8: Six Lowest Maximum Similarity Poems with Dr. Ling's Tokenization

Next, let's look at the six highest maximum similarities. These are poems numbered 3.19, 3.57, 3.43, 3.32, 3.24, and 3.9. The table showing the similarity metrics are shown below.

Poem Number	Name of Song	Maximum Similarity	Non-zero Sim Count
3.16	鷓鴣天	0.133	43
3.24	清平樂	0.133	27
3.27	殘梅	0.143	52
3.49	玉樓春	0.143	40
3.19	臨江仙	0.191	46
3.57	臨江仙	0.191	43

Table 9: Six Highest Maximum Similarity Poems with Dr. Ling's Tokenization

It is very interesting that out of the six lowest maximum similarities half of the poems are considered to be the most credible historically, two are problematic, and one is a less credible

poem. We would expect obvious misattributions to have the lowest maximum similarity value, so it is interesting that there are so many credible poems here.

We can also represent the distribution of these similarities graphically by using a histogram with the x-axis representing the maximum cosine similarity of a poem, and the y-axis representing the number of poems that fall into a group. The histogram is shown below.

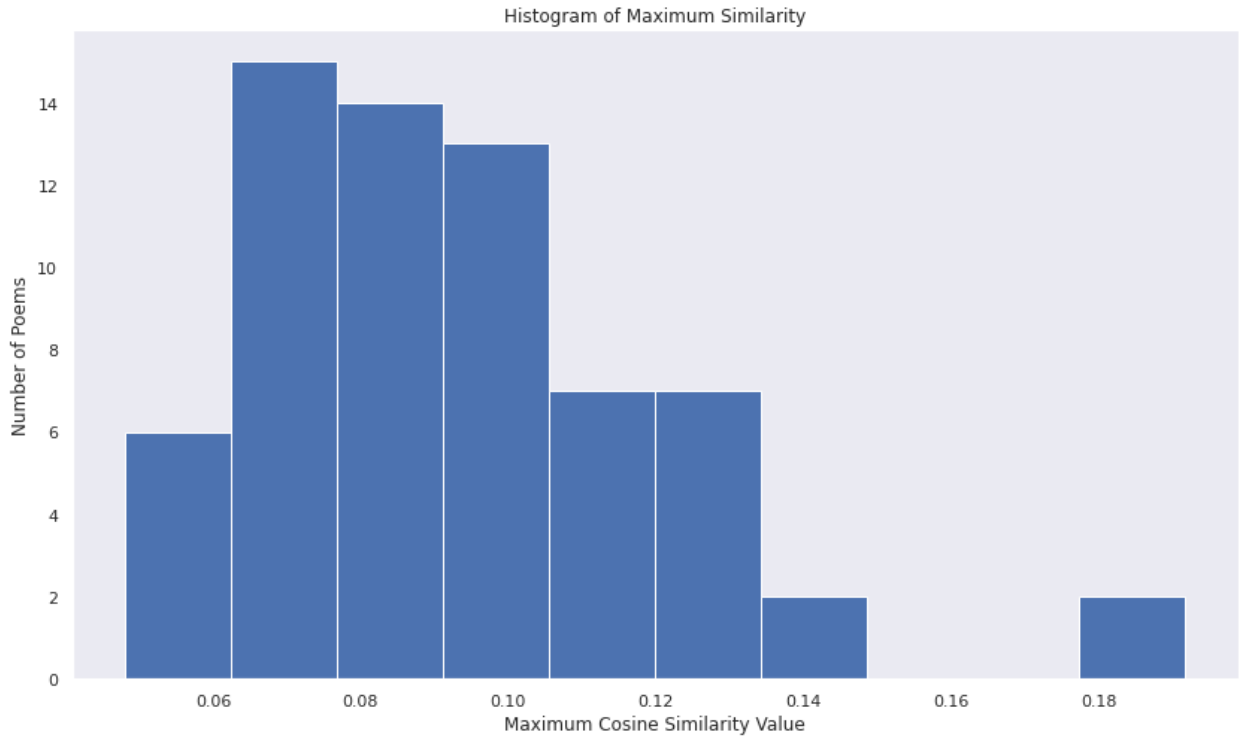

Figure 3: Histogram of Maximum Similarity Values with Dr. Ling's Tokenization

We can see that most of the poems have a maximum cosine similarity between 0.06 and 0.10. The poems that are located above 0.14 are poem number 3.19 and 3.57, the two poems discussed in chapter one which are based off Ouyang Xiu.

The cumulative similarities is also an important metric to look at. This describes the sum of all of the cosine similarities for a given poem. Thus, instead of just looking at the similarity of the poem that it matches the best, this metric measures how much similarity exists between this

poem and all other poems. When we use this metric, we see a different histogram as shown below.

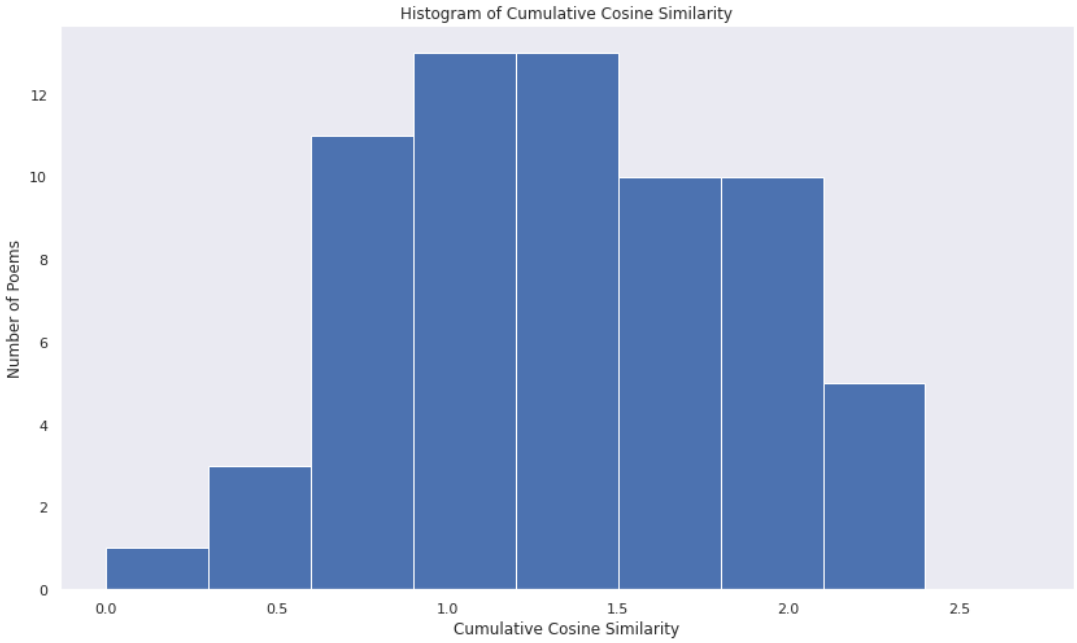

Figure 4: Histogram of Maximum Similarity Values with Dr. Ling's Tokenization

We see that a majority of the poems are between 0.7 and 2 for their cumulative similarities. The poems that fall outside of this region are of specific interest because they either have a small cumulative similarity, or a large cumulative similarity. These poems are of particular interest to us.

The ranking of cumulative similarity is different from the ranking of maximum similarity. This makes sense because there can be a poem which is very similar to another poem, but not many others. The six poems with the lowest cumulative cosine similarity are shown in the table below.

Poem Number	Name of Song	Max Sim	Non-Zero Sim Count	Cumulative Similarity

3.58	山花子	0.059	13	0.290
3.22	訴衷情	0.124	41	0.444
3.18	憶王孫	0.109	46	0.469
3.17	小重山	0.105	38	0.597
3.13	一翦梅	0.082	24	0.646
3.47	七娘子	0.112	21	0.647

Table 10: Six Lowest Cumulative Similarity Poems with Dr. Ling's Tokenization

The six poems with the highest cumulative similarity are shown below.

Poem Number	Name of Song	Max Sim	Non-Zero Sim Count	Cumulative Similarity
3.1	南歌子	0.073	21	2.077
3.48	憶少年	0.075	30	2.193
3.52	醜奴兒	0.075	22	2.265
3.31	元宵	0.082	44	2.283
3.28	紅梅	0.117	42	2.362
3.23	行香子	0.095	40	2.377

Table 11: Six Highest Cumulative Similarity Poems with Dr. Ling's Tokenization

This cumulative similarity metric and the maximum similarity metric helps us better understand these poems. The poems which have a high maximum similarity metric, but a lower cumulative similarity metric are poems which match the diction of one poem very well, but were less similar to the rest of the poems. An imitator may have written the poem based on one or two poems known to be authentic Li Qingzhao, resulting in a high maximum similarity metric, but a lower cumulative similarity metric (poems 3.57, 3.47, 3.22). On the other hand, an imitation where there are very few similarities would just be a poor imitation such as poem 3.58. The

poems that would be a better imitation would be the ones with more similarity such as 3.52, 3.48 and 3.31. Using this logic, we would also conclude that the poems which are historically credible and have a low cumulative similarity score are some of Li Qingzhao's more unique poems (3.3, 3.4, 3.5).

Experiment Using Character Tokenization

In this section, we repeat the comparison process using the TF-IDF cosine similarity function with the character tokenization technique. Since each character on its own is considered a token, there are many more similarities between the poems than in the previous method. There were a total of 1054 unique characters used within all of the 66 *ci* poems included within Ronald Egan's anthology. The most common characters used are shown below:

'雨': 20, '酒': 20, '深': 20, '情': 21, '天': 23, '玉': 23, '日': 23, '無': 23, '更': 23,

'梅': 25, '上': 26, '一': 28, '香': 32, '來': 36, '春': 38, '不': 39, '人': 44, '花': 46, '風': 48,

with the number corresponding to the frequency of the most common characters. It is clear that frequent characters like 風 (wind) and 花 (flower) are common topics to many poems of this style. There are also many characters such as 無 (without), 不 (not), and 更 (more) which are also very common. The usage of these more vernacular words is typical of Li Qingzhao's style and thus could be more easily imitated. Because there are so many common characters between all of the poems and more common words hold less significance, the TF-IDF model is a more appropriate process to compute cosine similarity. When we compare the poems, we get the similarity matrix below:

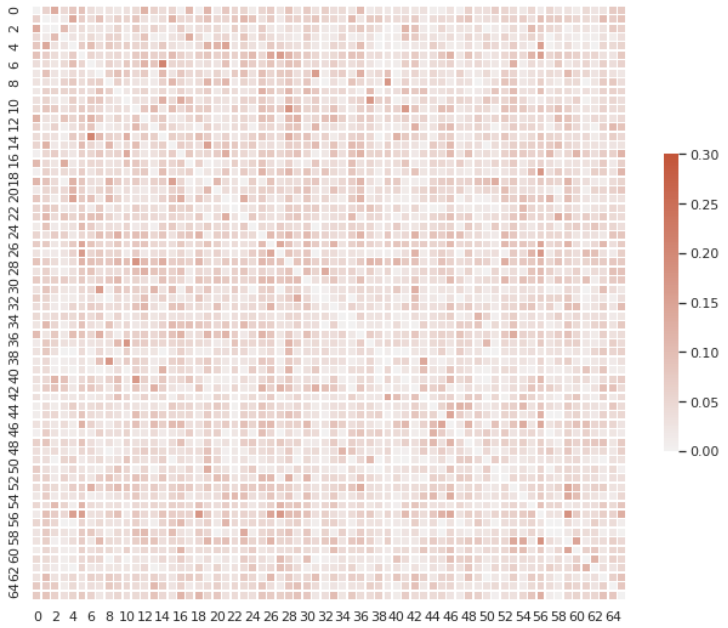

Figure 5: Heatmap of Cosine Similarity with Dr. Chao Ling's Tokenization

We see that there are more similarities than the regular term frequency bag of words using Dr. Chao Ling's tokenization. There are few poems that stand out. Looking at the cumulative cosine similarity, we see that the six poems with the lowest cumulative similarity for the character tokenization technique are shown below:

Poem Number	Name of Song	Max Sim	Cumulative Similarity
3.39	浣溪沙	0.107	2.043
3.51	點絳唇	0.109	2.193
3.52	醜奴兒	0.129	2.266
3.41	浣溪沙	0.121	2.309
3.4	如夢令	0.131	2.329
3.33	添字醜奴兒	0.134	2.413

Table 12: Six Lowest Cumulative Similarity Poems with Character Tokenization

Notice how poem 3.4 has some of the lowest similarity values for both methods (Dr. Ling's and character tokenization), meaning that this poem uses some of the most unique vocabulary within the anthology. We also see a huge change in the rank of poem 3.47. In Dr. Ling's tokenized method, it ranks in the bottom six, whereas in the character by character method, it ranks in the top six. Supposing 3.4 is authentic (which is convention) then this would be one of Li Qingzhao's more unique authentic poems. We also see that many of the poems that are more dissimilar are poems which come from later anthologies (3.51, 3.52, 3.41 and 3.39) and two of the poems are from earlier anthologies (3.33 and 3.4).

The six poems with the highest cumulative similarities are shown below.

Poem Number	Name of Song	Max Sim	Cumulative Similarity
3.47	七娘子	0.1433	4.0113
3.26	孤雁兒	0.1248	4.1380
3.6	咏白菊	0.1624	4.2422
3.27	殘梅	0.1470	4.3725
3.57	臨江仙	0.1821	4.6422
3.29	念奴嬌	0.1792	4.9527

Table 13: Six Highest Cumulative Similarity Poems with Character Tokenization

We see that poem 3.57 (poem that could be a replica of poem 3.19) has the second highest similarity score. Poems 3.26 and 3.27 both come from *Garden of the Plums*, which is the anthology that predates the *Elegant Lyrics for Music Bureau Songs*. It is interesting that these poems have some of the highest similarity of any of the other poems. Perhaps later poets used the language within these poems as a baseline for language included in the imitations. We also see the appearance of poem number 3.57 again. Not only is there a lot of similarity with poem

number 3.19, but we also see that many of the characters used within this poem are used within other poems in Li Qingzhao's anthology.

In conclusion, one of the differences between these two methods is the TF-IDF component which weighs the infrequent terms more heavily. The first method computed similarity for words and phrases (the way that the poems are intended to be read) and the second method looks at only the character usage. There was some overlap between the two experiments particularly poems number 3.4, 3.47, 3.57. We also see the presence of three of the five poems from *Garden of the Plums* as results in the experiments. This experiment highlights the vocabulary used within Li Qingzhao's poems. We can break down the poems within the Li Qingzhao anthology and better understand the words and characters that make up these poems and their respective frequencies. Using this method, we gain insight into which of the poems are lexically similar to each other and which poems are more unique. The hypothesis is that poems which have a higher cumulative frequency are either authentic and have had vocabulary borrowed from them or that they are imitations which relied heavily on the poems in previous anthologies. On the other hand, the poems which have low similarity could be poems which are unique Li Qingzhao poems, or poems which may have been misattributed to her, or poor imitations. In the next section, we will look at some of the poems which occur in the experimental results[6].

[6] A comprehensive list of the poems and their similarity rankings are provided in the appendix.

Chapter 3: Results and Authorship of Selected Poems

In this chapter, I discuss a number of typical poems based on the result of the experiments performed in Chapter Two. I divide them into four categories: 1. Historically authentic with low similarities (in other words, unique Li Qingzhao's poems); 2. Historically less credible with low similarities (poor imitation or misattributions); 3. Historically authentic with high similarities (authentic, typical Li Qingzhao's poems); 4. Historically less credible with high similarities (successful imitations or authentic). While there is no definitive answer on whether these poems are authentic Li Qingzhao or not, my hope is to provide some theories for the authenticity of certain poems backed by the cosine similarity results, the poem's histories, and a stylistic analysis of the poems.

The poems to be discussed are the following ones: lower similarity (3.3, 3.4, 3.5, 3.58, 3.64), higher similarity (3.23, 3.28, 3.49, 3.52), and a few curious poems (3.43, 3.57). The poems with lower similarity have very few terms that match the overall diction of all poems and the poems with higher similarity have many terms in common with other poems within the anthology. Because of this, we say that the poems which have lower similarity are more unique, and the poems with higher similarity are more typical. We can also differentiate the two different types of similarity metrics we use. The term maximum similarity is defined as the similarity between the particular poem and any other *single* poem. For example, if two poems are very similar, then we would say that they have a high maximum similarity. The term cumulative similarity is the metric which says how well a particular poem matches with *all* the other poems. The experiment that this chapter focuses on is Dr. Chao Ling's tokenization with term frequency bag of words.

Historically Credible Poems With Little Shared Vocabulary (Unique and Authentic)

In this section, the poems which have a small maximum similarity and a low cumulative similarity (3.3, 3.4, 3.5) are discussed. These are poems which are more unique and share few terms with other poems. Poems 3.3, 3.4, 3.5 are some of the earliest poems attributed to Li Qingzhao and are within the most credible group of poems. These three poems were first included in *Elegant Lyrics for Music Bureau Songs* 乐府雅词 in 1146. These are the more lexically unique poems in Li Qingzhao's anthology, with poems 3.3, 3.4, and 3.5 being ranked 61st, 66th, and 64th out of all the poems. These poems are highly reliable, yet unique. Why do some of the earliest poems have such little similarity with the other poems in Li Qingzhao's anthology? Wouldn't we expect imitators to use the dictions from previous poems to better disguise their poems as Li Qingzhao's? First, let's take a closer look at poem 3.3.

漁家傲	To the tune "The Fisherman Is Proud"
天接 1 雲濤 1 連 2 曉霧 1。	The sky joins billowing cloud-waves to morning mists.
星河 2 欲轉 1 千帆 1 舞 1。	The River of Stars begins to turn, a thousand sails dance.
彷彿 1 夢魂 3 歸帝所 1	My dreaming soul seems to have gone to the Lord of Heaven's place, where I hear Heaven speak.
聞 4 天語 1。	
殷勤 1問我 1歸何處 1。	What is your final destination, it asks, showing real concern.
我 2 報 2 路長 1 嗟 1日暮 2。	The road is long, I say, and the day already late.
學詩 1 謾有 1 驚人 1 句 2。	I write poetry, but my startling lines are produced in vain.
九萬里 1風鵬 1 正舉 1。	A wind blows thousands of miles, the giant phoenix will soon take flight.

風 12 休住 1。 蓬舟 1 吹取 1 三山 1 去 2。	Oh wind, do not slacken! Blow my little boat to the distant Isles of Immortals.(Egan 98-9)

Poem 3.3: To the tune "The Fisherman Is Proud" (漁家傲)

There are a few words in this *ci* poem that occur more than once in anthologies which include her works. We see that the term "Milky Way" (星河) appears in another poem in the *Elegant Lyrics for Music Bureau Songs*, poem 3.1. The term "dreaming soul" (夢魂) occurs in two later poems both from later anthologies, 3.65 and 3.66. The term "dusk" (日暮) occurs once more in poem 3.4 as well. We may expect poem 3.3 to have similarities to 3.1 and 3.4 because they are all from the same anthology and are historically authentic. If imitations, then poems 3.65 and 3.66 may have borrowed the usage of "dreaming soul" from poem 3.3 to create a more authentic style. Many other poems use the term "wind" (風) but this is a very common word.

Although relatively short, this poem includes three literary allusions. Li Qingzhao was very particular about the usage of literary allusions. In her short essay *On Song Lyrics* 詞論 outlining her opinions of the *ci* genre, she comments on how allusions should be used within the poems. Criticizing her contemporaries, she says, "Qin cares only about emotions and has too few literary references. 秦即專主情致，而少故實", thus criticizing not enough literary allusions in Qin Guan's *ci* poetry (Egan 58). She also attacks Huang Tingjian for the opposite, improperly using too many references. She says, "As for Huang, although he prizes literary allusions, his works have many defects. They are like jade that has blemishes, reducing its value by half. 黃即尚故實，而多疵病，譬如良玉有瑕，價自減半矣。" (Egan 58). It is clear that Li Qingzhao advocates for a style with a balanced use of literary and historical allusions, which is true to many of her authentic works.

The three literary allusions in poem 3.3 are Peng 鵬, 90,000 Li 九萬里, and the Three Mountains 三山. It is very interesting that she included Peng 鵬, a mythical bird, within her poem, because she never uses Peng in any of her other poems. However, in *A Letter Submitted to Hanlin Academician Qi Chongli* 投翰林學士綦密禮啟 [7], she references this story, saying:

> "The towering Peng bird soars high above, whereas the little quail sinks to the ground. The fire mouse and the ice silkworm can hardly share the same preferences. This is as obvious to little boys as it is to wise men.
>
> 高鵬尺鷃, 本異升沉;火鼠冰蠶, 難同嗜好。達人共悉, 童子皆知。" (Egan 66).

The story of Peng 鵬 and the phrase 九萬里 (90,000 Li) allude to Chapter One of *Zhuangzi* called "Easy Wandering逍遥游". The first few sentences of the chapter and the translation are provided below (Zhuangzi).

> "北冥有鱼，其名为鲲。鲲之大，不知其几千里也；化而为鸟，其名为鹏。鹏之背，不知其几千里也;怒而飞，其翼若垂天之云。"
>
> "In the northern sea is a big fish named Kun. This fish is huge, nobody knows how many thousands of Li long it is; The fish became a bird with the name of Peng. Nobody knows how many thousands of Li Peng's back is. When Peng flys, the wings are like clouds covering the Heaven." (Translated by me)

The term "90,000 Li" is also used extensively throughout the same *Zhuangzi* chapter. It is used to describe the huge distance that Peng ascends in the air and the distance Peng can travel. Zhaungzi's story includes a Cicada and Little Dove that are trying to understand Peng's experience, but realize they are unable to understand something of that magnitude. More specifically, they say, "I set my mind on leaping up and flying, jumping from the elm tree to the

[7] This letter talked about her remarriage, and her feelings towards herself.

sandalwood, but always end up not getting there, and fall back to the ground. How can he fly 90,000 li to the south? 我决起而飞，抢榆枋而止，时则不至，而控于地而已矣，奚以之九万里而南为？” (Muller). The small Cicada and Little Dove talk about some other things about Peng that they are incapable of understanding, leading them to acknowledge the idea that “Small understanding can't match great understanding; the short-lived cannot match the long-lived. 小知不及大知、小年不及大年。” (Muller). This is very similar to the reference within her letter, and can certainly help with understanding the poem’s meaning: .

The final reference we see is the three mountains (三山). In ancient tales, there are three mythical mountains in the middle of the East Sea which are hard for earthly people to reach. Poems often use this mythical, yet difficult and hard to reach place as a metaphor. They are mentioned in various texts, from the “Treatise of the offerings for Heaven and Earth”封禅书 in Sima Qian’s 司马迁 *Records of the Grand Historian* 史记 to Song Dynasty poems, such as Su Shi’s. Again, we do not see her using the term Three Mountains 三山 in other poems, but we do see other poems referencing the Three Mountains by their respective names, Penglai 蓬萊 (poem 3.14) and Jasper Terrace 瑤臺 (poem 3.44) (Egan 116 & 164).

Understanding the literary allusions can help us better understand the poem. For part of Li Qingzhao’s life, she would sail on boats from place to place. After the Jin armies attacked the capital, Li Qingzhao and her husband travelled south to Nanjing on a boat. In her *Afterword*, she says “When we reached Donghai, we crossed the Huai River in a string of boats. Then we crossed the Yangzi River and arrived at Jiankang 至東海，連艫渡淮，又渡江，至建康。” (Egan 76). It is possible that she is writing about her experience travelling south in poem 3.3. The poem talks of entering Heaven’s place where the Heavens genuinely asks her about her final destination. She responds that she already knows the road is long, and the day is late, where her

experience of writing poetry is in vain (perhaps she is referencing all of the poetry that she left in Qingzhou). Alluding to Peng, she asks the enormous bird to flap its wings and produce enough wind to propel her to the Three Mountains. There seems to be a general theme of understanding and hope. She hopes to end up in a distant and remote place that offers beauty, but understands that it requires the work of others to get her there. In the poem, she relies on the mythical Peng to flap its wings, and in her life, she is relying on others to safely get her to Nanjing. She invests her hope in Daoist transcendence by referring to Peng to overcome the difficulty in her life. She may also be talking metaphorically about achieving literary immortality. Since this poem is included in the most credible anthology, the usage of literary allusions matches her self-described style, my reading of the allusions shows that this was probably written when she first left for the south, and the number of similarities are low, we may conclude that this poem is a very unique and authentic poem by Li Qingzhao.

Another poem with very low similarity, yet a very credible history is poem 3.4. This poem has the lowest maximum similarity and is ranked 54th for cumulative similarity and it only shares diction seven times. Poem 3.4 is to the tune of "As If in a Dream 如夢令" and is also included in *Elegant Lyrics for Music Bureau Songs*. Again, it is curious that such an early and reliable poem has such a low similarity. The poem and Ronald Egan's translation are included below.

如夢令	To the tune "As If in a Dream"
常記 1 溪亭 1 日暮 2	I often recall one sunset in a riverside pavilion.
沈醉 3 不知 3 歸路 1	Having drunk too much, I forgot the way home.
興盡 1 晚 2 回舟 1	Knowing it was late, I started back in my boat at dusk

誤入 1 藕花 1 深處 1 爭渡 2 爭渡 2 驚起 1 一灘 1 鷗鷺 2	but paddled by mistake into a thick patch of lotuses. Struggling to get out, struggling to get out, I startled a whole sandbar of egrets into flight. (Egan 100-1)

Poem 3.4: To the tune "As If in a Dream" (如夢令**)**

Right away we see that there are no words that appear a significant number of times throughout the poems like we saw with wind (風) in poem 3.3. Instead, we see more terms that appear in just one or two other poems such as dusk (日暮) (poem 3.3), well drunk (沈醉) (poem 3.9 & 3.22), unknowingly (不知) (3.6 & 3.28), and egrets (鷗鷺) (3.18). All these repeated terms appear in poems that are considered reputable. Poem 3.4 only shares vocabulary with poems that also come from early and credible anthologies (aside from poem 3.28[8]). This means that imitations from later anthologies did not borrow words from poem 3.4.

Although there are only two other poems which contain the word "well drunk" (沈醉), there are many other poems which talk about drinking: The character alcohol (酒) and drunk (醉) occurs 20 and 12 times respectively throughout the poems. Aside from using well drunk (沈醉), we see the usage of "face flushed with wine" (醉臉) (poem 3.6), "drink and enjoy" (醉賞) (3.46), and "after drinking" (醉後) (3.63) all being used once throughout her anthology and the single character "drunk" (醉) being used 6 times (3.7, 3.16, 3.24, 3.25, 3.36 & 3.53). The term "alcohol" (酒) is used in several different contexts. The tokenized words that contain alcohol are as follows, and many of them only occur once: "sobering up" 酒醒 (poem 3.22 & 3.29), "after wine" 酒闌 (3.16 & 3.21), "lingering wine" 殘酒 (3.5 & 3.32). It is intriguing that these terms

[8] Although this is seen by Ronald Egan as less credible, it is still an early attribution to Li Qingzhao.

for alcohol are repeated within the more credible poems, yet the times where alcohol (酒) is used as a word, we see mostly in later poems (3.7, 3.36, 3.37, 3.55, & 3.60). We clearly see that Li Qingzhao recycles words from her other works.

Another difference we see between poem 3.4 and poem 3.3 is the use of literary allusions. In this poem, we do not see use of any well known literary reference. To summarize this poem, the author happily thinks back to a good time in her life where she got too drunk while boating and ended up in a deep patch of lotuses. She then struggles to get out, scaring away birds. The scenery in this poem is beautiful, simple, and precise. This poem is one of the more unique authentic poems by Li Qingzhao. Perhaps one reason this poem has low similarity with other poems is because it is a happier poem unlike many of the other poems, and thus one of her earlier works, before the Jurchen invasion.

Just like poem 3.4, poem 3.5 is to the tune "As If in a Dream" 如夢令 and was included in the same anthology.

如夢令 昨夜 2 雨疏 1 風驟 1 濃睡 1 不消 1 殘酒 2 試問 1 捲簾 1 人 20 却道 1 海棠 2 依舊 1 知否 2 知否 2 應是 3 綠肥 1 紅瘦 1	To the tune "As If in a Dream" Last night the rain was intermittent, the wind blustery. Deep sleep did not dispel the lingering wine. I tried asking the maid raising the blinds, who said the crab-apple blossoms were as before. "Don't you know? Don't you know? The greens must be plump and the reds spindly." (Egan 100-1)

Poem 3.5: To the tune "As If in a Dream" (如夢令)

Right away we see that the word person (人) (maid in this case) is used 20 other times throughout the anthology. Aside from this common word, all the other words only occur a few other times throughout the anthology. We see that "last night" (昨夜) also occurs in poem 3.45, "lingering alcohol" (殘酒) occurs in poem 3.32, "crab-apple" (海棠) occurs in poem 3.21 and "should be" (應是) occurs twice more in later poems (3.46 & 3.57)[9]. We also find some similarities between poem 3.4 and 3.5 that extends beyond diction. For example, they both discuss failing memory while being surrounded by natural objects. In poem 3.4, the author is drunk and unable to remember the way home. In poem 3.5 the poet is still drunk in the morning and is unaware of the conditions of the crab-apple blossoms outside. These topical similarities between the poems were undetected by our algorithm, yet they still play a role in human understanding of the poem's similarity. Human readers are able to pick up these similar themes throughout poems that simple analyzing of diction cannot.

Less Credible Poems With Little Shared Diction (Misattribution or Poor Imitation)

Now, let's talk about the poems with historically less credible Li Qingzhao attributions that have low cosine similarity (3.58, 3.64). Poem 3.58 is one of the most dissimilar poems in Egan's translation and it's first appearance was in 1583's *Huacao Cuibian*. According to our algorithm, poem 3.58 ranks last in cumulative similarity and 63rd out of 66 poems for maximum similarity. Below we can see that there are only two poems that have less than 15 non-zero cosine similarities, poem 3.4 and 3.58 (located all the way to the left). Poem 3.58 has common language with thirteen other poems and is the second bar from the y-axis, clearly smaller than the rest of the distribution.

[9] Both of these poems were included in *Mei Yuan* without attribution.

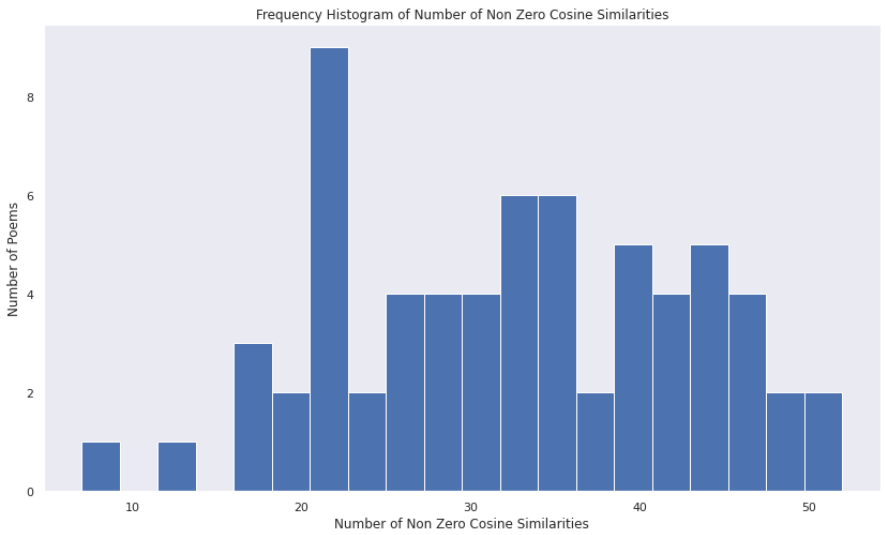

Figure 6: Histogram of Number of Non Zero Cosine Similarities

All poems which have similarity with 3.58 were included in the *Huacao Cuibian* except for poem 3.65. A majority of these similarities were just based off two words: "light" (輕) and "dream" (夢). Below is a histogram of the non-zero magnitudes of the cosine similarities[10]. Many other poems extend into the 0.08 to 0.12 area, but this poem has none[11].

[10] When compared with similar histograms from other poems, this graph has one of the fewest and smallest distributions of cosine similarities.
[11]The histogram for each poem's cosine similarity is included in the appendix for comparison.

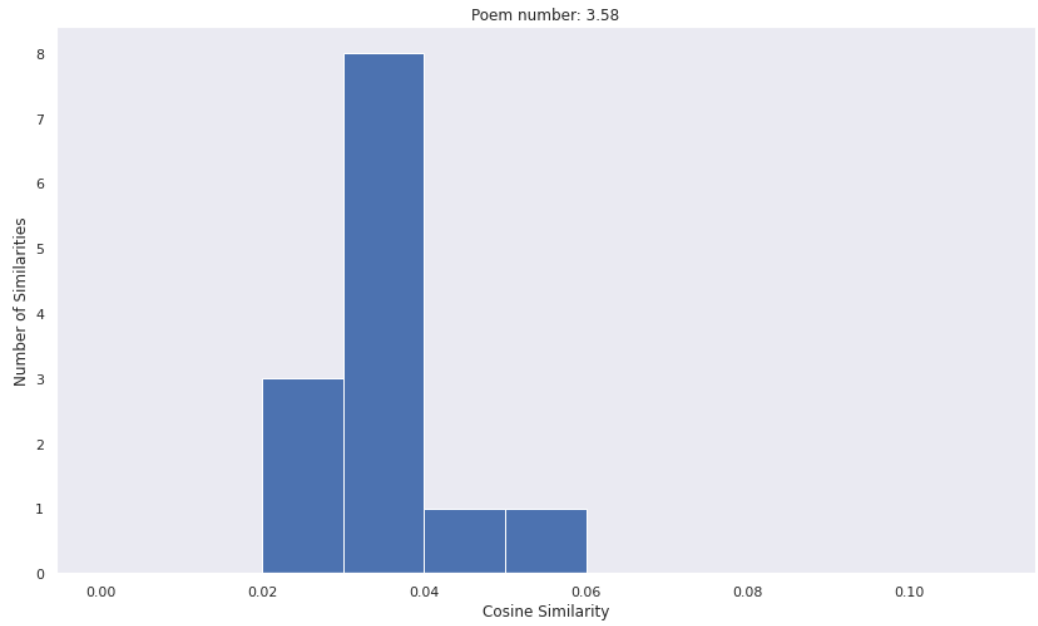

Figure 7: Histogram of Poem 3.58's Cosine Similarity Values

Because of these reasons, this poem was flagged as suspicious. Why would this poem have such low similarity to the other poems? Since this poem is historically not credible, what implications does a low cosine similarity have on its authenticity? And where do most of the similarities come from? We must read this poem and understand the background of the similar poems before we can attempt to answer these questions. Below we have the poem with the number of times each term is present throughout the entire anthology.

山花子	To the tune "Wildflower Seeds"
揉破 1 黃金 1 萬點 1 輕 5 剪成 1 碧玉 1 葉 1 層層 1 風度 1 精神 2 如 1 彥輔 1 太 1 鮮明 1	Yellow gold, as if torn apart, into myriad dots of blossoms. Green jade, shaped with scissors, the multiple layers of leaves. In style and spirit like Yanfu, so fresh and bright! Plum blossoms, row after row, how vulgar they are!

梅蕊 2 重重 1 何 1 俗 1甚 2 丁香 1 千結 1 苦 2 粗生 1 熏透 1 愁人 2 千里 3 夢 4 卻 1 無情 2	The thousand-petal lilac is crude by comparison. But your perfume wakes the dreamer from her distant journey home— how could you be so heartless! (Egan 184-7)

Poem 3.58: To the tune "Wildflower Seeds" (山花子**)**

There are thirteen poems[12] which have common words with poem 3.58 and there is a higher similarity from poem 3.29 and 3.64 which both share two similarities with the words "light" (輕), "thousand li" (千里), "worried person" (愁人) and "dream" (夢). The rest of the diction in this poem rarely occur throughout the rest of the anthology. Because poem 3.58 never appeared in any other previous anthology, and a majority[13] of the common words are included within poems in *Huacao Cuibian*, it is entirely possible that this poem is an imitation of Li Qingzhao's work that relied on the diction of the other Li Qingzhao poems in the *Huacao Cuibian*.

We also see a literary reference in this poem with the metaphor comparing the style and spirit to Yanfu. Ronald Egan says, "These lines paraphrase praise of the unworldly qualities of Le Yanfu (Le Guang 樂廣, d. 304) found in his official biography in the *History of the Jin* 晋史." (Egan 185). The *History of the Jin* would be familiar to Li Qingzhao, a learned poet. This conforms to Li Qingzhao's practice of alluding to literary references. However, since this poem was first attributed to Li Qingzhao in 1583, and there are very few shared diction with other poems, this poem is likely a less convincing imitation. The author probably used other poems

[12] Poems with common words are 3.6, 3.8, 3.16, 3.22, 3.23, 3.29, 3.34, 3.57, 3.29, 3.50, 3.55, 3.64, and 3.65.
[13] Only poems 3.65, 6.33, and 3.50 are not included in the *Huacao Cuibian* 花草粹编.

attributed to Li Qingzhao within the *Huacao Cuibian* as inspiration for this poem. If not an imitation, then it may just be a misattribution from another poet which publishers tried to pass as Li Qingzhao's. This is especially likely because the *Huacao Cuibian* anthology appeared around 400 years after Li Qingzhao's death and it attributed nine new poems to Li Qingzhao for the first time (Egan 96-97).

The last poem in this section is poem 3.64. This poem is to the tune of "Rankings" 品令[14] and ranks 62th and 51st for maximum and cumulative cosine similarity. However, it does have similar words with twenty-eight other poems.

品令 零落 1 殘紅 2 恰 2 渾似 1 胭脂 1 色 2 一年 2 春事 2 柳 3 飛 1 輕 5 絮 1 筍 1 添 3 新竹 1 寂寞 4 幽閨 1 坐對 1 小園 1 嫩綠 1 登臨 1 未 7 足 1 悵 2 遊子 1 歸期 2 促 1 他年 1 清夢 2 千里 3 猶 3 到 8 城陰 1 溪曲 1 應有 1 凌波 1 時為 1 故人 1 留目 1	To the tune of "Rankings" Faded reds lie scattered, looking just like the rouge on her face. What's left of this year's spring? The willows' gauzy fluff has gone flying, shoots have formed into small bamboos. Lonely now in the women's quarters, she sits facing the small garden's tender green. It's not enough to climb high to look out— she longs for the one traveling far away, hoping the day of his return comes quickly. Or perhaps some future day her clear dream will cross a thousand miles even to a hidden place beside the wall, a bend in the stream, Where she, like the goddess who trod on waves, will attract her lover's fixed gaze. (Egan 194-7)

Poem 3.64: To the tune of "Rankings" (品令)

[14] Master Ou Yangxiu 欧阳修 also wrote a song lyric to this song.

Although this poem was first attributed to Li Qingzhao in 1583 in *Huacao Cuibian*, it was actually present in the most credible anthology, *Elegant Lyrics for Music Bureau Songs*, and was attributed to another poet. Since this poem was previously recorded, it rules out the possibility of an imitation, and instead opens up the possibility of a misattribution by the editor of the *Huacao Cuibian*. Since the maximum similarity is so low, but the cumulative similarity is ranked slightly higher, it is possible that the publishers of *Huacao Cuibian* decided that this poem was similar enough to the other poems attributed to Li Qingzhao, and thus attributed it to her to draw attention.

We see topics of loneliness, dreams and travelling in this poem. These topics are similar to other poems in our anthology and some may think that it passes as an authentic Li Qingzhao's poem. This poem describes a longing for a man to return. Perhaps it is describing Li Qingzhao's longing for her husband, however it feels more political than that. The second half of the poem feels less convincing and lines 9–14 almost feels like it was written by a man. This is a good poem and it embodies some of Li Qingzhao's emotions, but it is likely not Li Qingzhao. Based on the low cosine similarity and the historical background of having a different attribution, it is likely that this poem is the work of another poet.

Historically More Credible with More Shared Vocabulary (Authentic and Typical)

Now that we have looked at a few poems that use little vocabulary common to the rest of the anthology and discussed the implications behind those, let's look at some poems that have some of the most shared terms. We will be focusing on poems 3.23 and 3.28 which have some of the higher maximum similarity and cumulative similarity ranks. These are poems that share many terms with other poems and they represent the typical style of Li Qingzhao. The first poem

discussed in this section is poem 3.23 and was included in the most credible anthology, *Elegant Lyrics for Music Bureau Songs*. It has the highest cumulative cosine similarity rank and its maximum similarity score is ranked 30th. We must ask ourselves why this poem has such a high cosine similarity. Does this poem touch upon a lot of topics also rendered in other poems throughout the anthology? Do other poems use the language in this poem?

Let's first look at the distribution of cumulative cosine similarities for the poems in Ronald Egan's anthology. The histogram below shows the number of poems that are at each level of cumulative cosine similarity. Poem 3.23 is located all the way to the right along with four other poems, just below the 2.5 value.

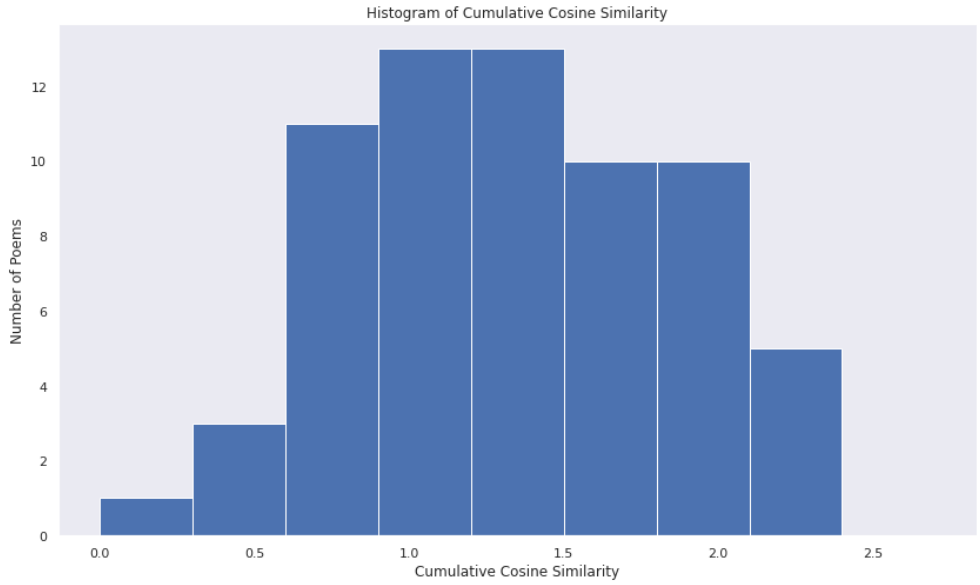

Figure 8: Histogram of Cumulative Cosine Similarity for Poems

This distribution shows that a majority of the poems have a cumulative similarity score between 0.5 and 2, with only eight poems falling outside of that range. Since poem 3.23 has the highest cumulative cosine similarity score, let's look at a histogram of its cosine similarities below.

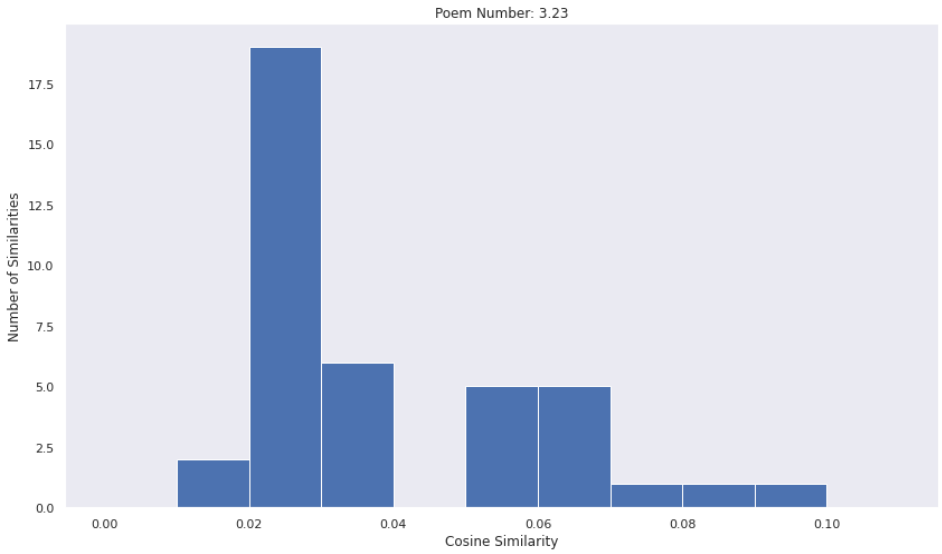

Figure 9: Histogram of Poem 3.23 Cosine Similarity Values

Looking at the distribution of cosine similarities for this poem, we can see that there are more similarities, with much higher values than for poem 3.58 in the previous section. We see that there are two poems which match it relatively well (3.66 and 3.51). Let's now read the poem to explore why this poem matches others so well.

This poem is to the tune of "Incense Offering" 行香子 and the poem, word frequency, and the English translation are provided below.

行香子 草際 1 鳴蛩 1 驚 2 落 3 梧桐 5 正 3 人間 3 天上 3 愁濃 1 雲階 1 月色 1 關 1 鎖 1 千重 1 縱 2 浮槎 2 來 10 浮槎 2 去 2 不 8 相逢 2 星橋 1 鵲駕 1	To the tune of "Incense Offering" Chirping crickets in the brush startle paulownia leaves off the branch. This is a time of deep sadness, in the heavens as on earth. A stairway of clouds to a moonlight terrain, the thousand gates are locked shut. Even if the raft comes floating by, it drifts on by, and never encounters Herd Boy.

經年 1 才 2 見 3 想 1 离情 1 别恨 1 難 4 窮 1 牽牛 1 織女 1 莫 8 是 3离中 1 甚 2 霎儿 3 晴 2 霎儿 3 雨 6 霎儿 3 風 12	A bridge across the River of Stars, formed by magpies meeting only once a year. Their parting sorrows and regrets never come to an end. Herd Boy and Weaving Maid— all they know is separation. Truly, theirs is a moment of clear sky, a moment of rain, a moment of wind. (Egan 130-132)

Poem 3.23: To the tune of "Incense Offering" (行香子)

Since this poem has diction in common with 40 of the 66 poems, we see that the frequency of words is much higher in this poem compared to the earlier poems. Some of these words are traditionally more common words like come (來), wind (風), or do not (莫), but we also see many words which are used in two or three different poems. This is much different than the sparse similarities we saw in the previous section. A big reason for the large cosine similarity is the use of shorter, more vernacular words such as come, none, who and popular imageries like wind, rain, and paulownia. These words seem to be very common to many of the poems thus giving this poem a large cumulative cosine similarity. Some researchers choose to eliminate words that are common to treat these words as stop words and remove them. However, the use of these words is important within the poetry because each character is thoughtfully chosen by the author.

We also see the use of a literary allusion within this poem. The twelfth line reads: Herd Boy and Weaving Maid 牽牛織女. The Herd Boy represents Altair and the Weaver Girl symbolizes Vega and they are separated by a river (the Milky Way). Once a year on the seventh day of the seventh month, a flock of magpies form a bridge so they can reunite. This story is very well known throughout Chinese folklore ("牵牛织女"). In poem 3.23, the author describes a very sad scene filled with loneliness and neverending separation. This poem uses the imagery of

the past story of the Herd Boy and Weaving Maid very well, and this may be a description of Li Qingzhao's life when Zhao Mingcheng abandoned her in the river, or afterwards when she was alone travelling from place to place.

Since this poem is credited to Li Qingzhao in *Elegant Lyrics for Music Bureau Songs*, it uses literary allusions in an easily understandable way, and there is a high level of similarity between this poem and many other poems, it is highly likely to be an authentic Li Qingzhao poem. Perhaps the poem's similarity is so high because it is a poem about sadness and loneliness, a recurring theme throughout the works attributed to Li Qingzhao. Alternatively, perhaps other poets used the diction in poem 3.23 to help shape their imitations and published them in later anthologies.

Poem 3.28 is to the tune of Spring in the Jade Tower 玉樓春 and the title is "On the Red Plum" 紅梅. This poem is very interesting because in 1129 it was first published in *Garden of the Plums* by Huang Dayu. Although this anthology is very early and contains six poems attributed to Li Qingzhao, there is a clear misattribution of one of the poems[15]. Because of this misattribution, Ronald Egan considers the five remaining poems to be less credible historically.

Although this poem is less credible according to Egan, it ranks 2nd for cumulative similarity and 13th for maximum similarity. In other words, this poem contains some of the most common diction throughout the entire anthology. With 42 poems containing similar diction and a few poems with stronger similarity, the question of authenticity arises. Could this be Li Qingzhao's work? How is it similar to other poems credited to Li Qingzhao? Let's read the poem to find out. Below we see the tokenized poem with their respective word frequencies as well as Egan's translation.

[15] One of the poems shows up in an earlier anthology of Zhou Bangyan (Egan 93).

紅梅 紅酥 1 肯 1 放 1 瓊苞 1 碎 1 探著 1 南枝 3 開 5 遍 2 未 7 不知 3 醞藉 2 幾多 1 時 3 但見 1 包藏 1 無限 4 意 3 道人 1 憔悴 8 春窗 1 底 1 悶損 1 闌干 4 愁 7 不 8 倚 4 要來 1 小看 1 便 2 來 10 休 3 未必 1 明朝 1 風 12 不 8 起 4	To the tune “On the Red Plum” The red cream displays itself, the jeweled pod bursts. I look to see if the southern branch is fully in blossom yet. There’s no telling how long they were concealed in preparation, all we see is the boundless feeling they contain. The blossoms know the person beside the spring window is haggard, troubled beside the balcony railing, too sad to lean and gaze afar. If you want to come view them a while, please do! There’s no guarantee wind will not rise in the morning. (Egan 139-41)

Poem 3.28: to the tune “On the Red Plum” (紅梅**)**

If this poem only showed up for the first time 300 years after Li Qingzhao’s death then it would be easy to speculate that it was an imitation, potentially very successful, since it uses so much language found in other poems in her anthology. However, since this poem actually predates almost all of the other poems attributed to Li Qingzhao, estimation of authenticity is more complicated.

We see the terms “railing” (闌干), “worry” (愁) , “not know” (不知), “come” (來), “wind” (風) and “haggard” (憔悴). These terms, along with many other lower-frequency terms, show up throughout many different poems in the anthology. The poem depicts the mood of sadness and it contains lots of vernacular language. There is common diction to sixteen of the twenty-four poems in *Elegant Lyrics for Music Bureau Songs*, and has the maximum similarity value with poem 3.6.

Because of these reasons, and the low chances of this being an imitation, there is no compelling reason to think that this would not be Li Qingzhao's work. This conclusion is based purely on diction and historical credibility. (The only other possibility is that editors of the anthology, intentionally or not, misattributed the poem that has some of the most common diction Li Qingzhao used as Li Qingzhao's.)

Less Credible with High Similarity (Authentic or Good Imitation/Misattribution)

In this section I talk about poems which are historically less credible but have a lot of shared vocabulary with other poems attributed to Li Qingzhao. In other words, they might be authentic, but they are more likely to be good imitations or misattributions. Another poem which is concerning, for different reasons, is poem 3.49. This poem is ranked third for maximum similarity and 48th for cumulative similarity. It was first included without attribution in *Garden of the Plums* in the twelfth century and was attributed to Li Qingzhao about 250 years later in the *Yongle Great Encyclopedia*. Since this poem was first published in Li Qingzhao's time, it is more likely a poem by another author which happens to share a lot of similarity to some of her other poems than an intentional imitation of her work. Perhaps publishers read this poem and saw how similar it is to other poems within Li Qingzhao's anthologies and attributed it to her. The poem is provided below.

玉樓春 臘前 1 先 2 報 2 東君 6 信 5 清 1 似 6 龍涎 1 香 6 得 2 潤 1	To the tune "Spring in the Jade Tower" Before the twelfth month the Lord of the East's messenger arrives. It's as pure as dragon nectar incense, fragrant and moist. The pale yellow blossoms refuse to open all at

黃輕 1 不肯 1 整齊 1開 5 比著 1 江梅 4 仍 1 更 11 韻 3	once— compared to the river plum, it is more elegant by far.
纖枝 1 瘦綠 1 天生 1 嫩 1 可惜 1 輕寒 1 摧 1 挫 1 損 5 劉郎 1 只解 1誤 1 桃花 2 惆悵 1 今年 4 春 11 又 8 盡 6	The slender branches, thin and green, are naturally delicate. Too bad they are readily harmed by even slight cold. Master Liu only knew to be deluded by peach blossoms, how sad that this year's springtime is already spent (Egan 172-3)

Poem 3.49: To the tune "Spring in the Jade Tower" (玉樓春**)**

The poem references the Lord of the East 東君, a main character in one of Qu Yuan's poems[16]. Five other poems[17] in Li Qingzhao's anthology also reference to the Lord of the East. Some of the words in this piece are also shared among others, however, it only shares a limited number of words with a handful of poems. The reason this poem has such high maximum similarity is the common words[18] with poem 3.27. Because of this, it is ranked high for maximum similarity, but fairly average for cumulative similarity. It is possible that this poem is from a different author, and the publishers felt that it matched similarly enough to Li Qingzhao's other poems to be credited to her. It is likely not an imitation because of the time period of its production and the low maximum similarity. If not authentic, it is likely a misattribution, in which case it was written by a skilled, but less famous poet since it shares some similarities based on diction.

[16] *Nine Songs* 九歌.
[17] The five poems which reference Lord of the East: 3.17, 3.37, 3.46, 3.47, 3.61
[18] The common words are: 信, 江梅, 更, 韻, 春, 又, 盡

The final poem in this section is poem 3.52. Poem 3.52 ranks 4th for cumulative similarity and 48th for maximum similarity. The first time this poem appears in any anthology is in 1550, in *Cilin Wanxuan*. Since it first appears long after Li Qingzhao's death, it is a possible imitation or misattribution. It only shares similar diction with one third of the poems.

醜奴兒 晚來 4 一陣 1 風 12 兼 1 雨 6 洗盡 1 炎光 1 理罷 1 笙簧 1 却對 1菱花 1 淡淡 2 妝 1 絳綃 1 縷薄 1 冰肌 2 瑩 1 雪 2 膩 2 酥香 1 笑語 2 檀郎 1 今夜 1 紗廚 1 枕簟 2 涼 3	To the tune "The Vile Charmer" This evening a storm of wind and rain washed away the blazing heat. Having finished playing the flute, facing a caltrop mirror she lightly dabs on makeup. Beneath purple thin silk her ice-like skin glimmers, luster of snow, milky and fragrant. Smiling, she tells her beloved, "Tonight, the mat and pillow behind the gauze bed-curtain should be cool." (Egan 178–9)

Poem 3.52: To the tune "The Vile Charmer" 醜奴兒

This poem also uses natural elements like rain and wind, but in a positive way since it washes away the blazing heat. At the end of the poem, the author again ties the theme of temperature while anticipating the night with her lover. This is very playful, seemingly flirtatious, and probably occurred during the beginning of her relationship with Zhao Mingcheng. The time period is likely before she had to relocate out of Qingzhou. If this poem is an imitation, then it would be an imitation of one of her earlier works which is less common. It could also be a clever manipulation of Li's poetic diction to create an erotic implication so as to meet the need of

the commercial readership in the late Ming dynasty. In the future it would be interesting to see which poems are about love, and perhaps create a timeline of when each poem likely was written in Li Qingzhao's life.

Other Selected Poems

In this section, Poem 3.43 and 3.57 are discussed. Although poem 3.43 is not at the extreme ends of the cosine similarity distribution, it still has interesting qualities that were flagged by the algorithm. Poem 3.57 which is discussed in Chapter One is also revisited with a computational lens.

Poem 3.43 is likely to be a really good imitation or a misattribution since it draws from the language used in Li Qingzhao's earliest and most available anthologies. The first time that poem 3.43 was recorded in any anthology was in the 1300s, some 150 years after Li Qingzhao's death. It uses similar language to thirty-three of the sixty-six poems, and five poems have their highest match with them. However, the cumulative similarity for this poem is ranked 50th out of sixty-six. It blends in really well with the rest of the poems in terms of cumulative similarity as well as maximum similarity. The reason that I think this is an imitation is because it first appeared in *Mao Jin's edition of Jade for Rinsing the Mouth* 毛晋本漱玉词 and it shares the maximum similarity with five poems from the more credible anthologies. The author likely read Li Qingzhao's most credible poems, and used the diction within them to fabricate a poem which matched them all very well.

點絳唇 寂寞 4 深閨 1	To the tune "Dabbing Crimson Lips" Lonely, deep in the women's quarters,

柔腸 1一寸 1 愁 7 千縷 1。 惜春 1 春去 1。 幾點 1 催 5 花雨 1。 倚遍 1 闌干 4 祇是 1 無 8 情緒 1。 人 20 何處 4。 連天 1 芳樹 1 望斷 1 歸來 4 路 2。	every inch of fragile innards has a thousand threads of sorrow. I cherish spring, but spring departs. Drops of rain hasten the blossoms. Having leaned everywhere on the balcony's railing, I have no enthusiasm for anything. Where is that person now? Fragrant trees stretch to the horizon, I gaze to the end of the road back home. (Egan 162-5)

Poem 3.43: To the tune "Dabbing Crimson Lips" (點絳唇**)**

Here we see the usage of many similar terms to other poems. We see that "lonely" (寂寞) (3.2, 3.32, 3.64) shows up 3 times other times. Perhaps Li Qingzhao's poems which showed sad emotion were more sought after, and thus had more imitations. The term "worries" (愁) and "without" (無) shows up frequently as well and is very typical of a Li Qingzhao poem. The term "railing" (闌干) shows up four times throughout her collective anthology. The other poems which use this are 3.28, 3.29, and 3.46. This imagery may have stood out to someone who wanted to write in the style associated with Li Qingzhao, and thus recycled the term.

We also see some influence from Yan Shu, another *ci* poet who lived shortly before Li Qingzhao and is often grouped with Li Qingzhao in the school of "delicate-restraint" (*wanyue*婉约). In his poem *Butterfly Loves a Flower* 蝶恋花, we see some lines similar to the second stanza of poem 3.43 in terms of imagery (railings and plants) and emotion (sadness). The poem describes the author who "climbs a high tower alone, gazing towards the end of the roads to sky" 独上高楼，望尽天涯路 and opens up with "chrysanthemums by the railings, melancholy smoke, the orchids cry out dewdrops"槛菊愁烟兰泣露. The context behind this line is that the author wants to send their special person a letter, but there are many obstacles in the way and they do not know where exactly they are. This is very similar to 3.43 where the author asks "Where is

that person now?人何處 .” “*Wangjin*望尽” is almost identical with “*wangduan*望断” as in “I gaze to the end of the road back home望斷歸來路”. Since this poem first appeared 150 years after Li Qingzhao’s death, it is possible that this is a misattribution of Yan Shu, or an imitation that drew language from both Li Qingzhao and Yan Shu.

This poem and the comparison with Yan Shu’s piece also raise an intriguing question. Between authors within the same school, how can we differentiate and tell them apart? Similar to painting practices, when one writes/paints in the style of an iconic author, is the final product copy, imitation, forgery or authentic but less original creation? It is exactly because of the existence of such nuanced definition of authorship/authenticity that the pre-modern editors of poetic anthologies felt free and comfortable to attribute many poems, even just *in style of* Li Qingzhao (or her fellow *wanyue* poets, eg., Yan Shu), to be her work.

The last poem in this chapter is one that we have discussed previously in Chapter One, poem 3.57. This poem ranks first for maximum similarity and 17th for cumulative similarity. We can see how this maximum similarity compares to others by looking at a histogram of the maximum cosine similarity values.

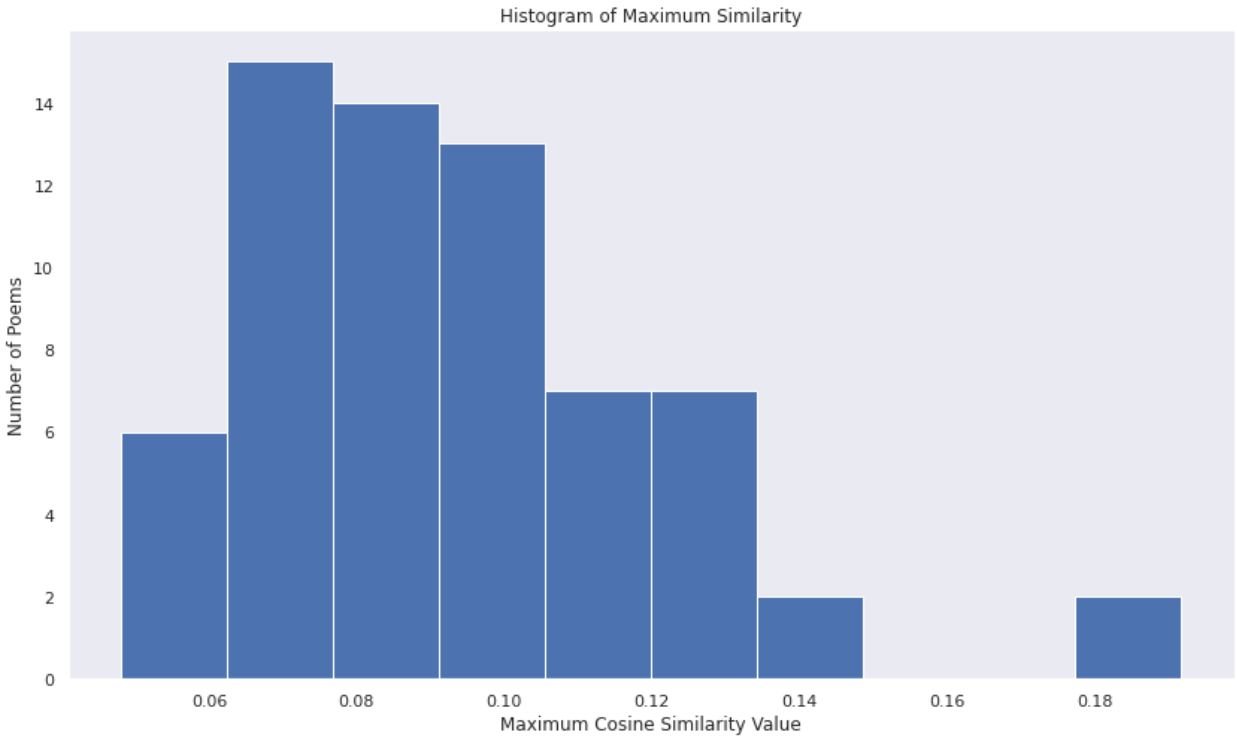

Figure 10: Histogram of Maximum Cosine Similarity Values

We see that poem 3.19 and poem 3.57 are the only poems in the farthest bin to the right. These definitely stand out for having such a large maximum similarity value. Before computing the cosine similarity values, we had already expected this poem to be of interest since a lot of the diction matched poem 3.19, the other poem that borrowed language from Ouyang Xiu. These two poems share the first two lines, so it makes sense that our algorithm placed poem 3.57 as having the largest maximum similarity.

Poem 3.19 is considered to be in the most credible category because it is present in *Elegant Lyrics for Music Bureau Songs* 乐府雅词, yet it's cumulative similarity is lower than poem 3.57 which is less historically credible. Poem 3.57 and the term frequencies are provided below, along with the translation.

臨江仙 庭院 6 深深 2 深 9 幾許 3 雲窗 2 霧閣 2 春遲 1 為誰 1 憔悴 8 損 5 芳姿 2 夜來 5 清夢 2 好 10 應是 3 發 1 南枝 3 玉 4 瘦 3 檀 1 輕 5 無限 4 恨 7 南樓 1 羌管 2 休吹 1 濃香 1 吹盡 1 有誰 1 知 3 暖風 2 遲日 1 也 6 別到 1 杏花 1肥 1	To the tune “Immortal By the River” Deep the deep courtyard, how deep is it? Cloudy windows and misty halls, late in spring. For whom are you so weakened, your fragrant beauty diminished? Last night in my lovely dream you were fine, I thought you’d be filling the southern branches. The jade is grown frail, the sandalwood hue faded, how sad! Don’t let the Tibetan flute play its melody in the southern loft. When your fragrance is blown away who will know? The wind is warm, the days of sunshine long, and the apricot blossoms plump. (Egan 184-5)

Poem 3.57: To the tune “Immortal By the River” (臨江仙**)**

We see a lot of similarity not only in the first line, but also scattered throughout the rest of the poem. The poet used many terms which are similar to other Li Qingzhao poems such as “haggard” (憔悴), “loss” (損), “last night” (夜來), etc. This poem contains the use of many vernacular terms common to other Li Qingzhao’s attributed poems. If this poem was to have first appeared in the *Huacao Cuibian*, then it would make sense to think of this as an imitation, given that the first line was provided in Li Qingzhao’s annotation of 3.19. It would have been quite easy for someone to construct a poem with the first line and diction from other Li Qingzhao attributions. However, this poem actually occurs for the first time in 1129 in *Garden of the Plums*, predating it’s most similar poem, 3.19. In this anthology, the poem is without attribution.

Could it be that the publisher included an authentic Li Qingzhao poem without attribution? Since poem 3.57 was published before poem 3.19 (which had the annotation describing the structure), I find it unlikely that it was an imitation. Instead, this is either an authentic Li Qingzhao poem, or another poet also enjoyed including the same line from Ouyang Xiu's poem and wrote using similar diction to the rest of the anthology.

Conclusion

In this thesis, I examine the authorship in Li Qingzhao's anthology through a computational lens. This experiment is the first analysis of Li Qingzhao's diction to help determine authorship with the help of natural language processing. Using cosine similarity with bag of words, I show that the language used within the corpus of poems can help identify poems with unique diction, and those which share diction with other poems in the anthology. After ranking these poems by various similarity metrics, I investigate specific poems, their histories, and their literary contents with the goal of establishing authenticity. Specifically, I found some results that agree and some that disagree with literary historians' conclusions. I also offer speculation on the type of misattribution and discuss the quality of the imitation. Out of the eleven poems that I examined more closely, I found that poems 3.3, 3.4, 3.5, 3.23, 3.28, and 3.57 are likely to be authentic Li Qingzhao. On the other hand, I found that 3.58 seems to be a less convincing imitation and poem 3.49 and 3.64 are likely to have been a misattribution from another poet if not authentic. Poem 3.43 is likely a convincing imitation, and poem 3.52 seems to be a mediocre imitation of her earlier works.

Some of these results agree with Ronald Egan, whose analysis was based purely off of historical reasons. In Egan's book, the *Burden of Female Talent*, the poems are separated into four different levels of credibility with one being most credible, and four being least credible. Due to my findings, I suggest that poem 3.57 be moved from the least credible category to the second most credible category. I also believe that poem 3.28 should be moved from the third category to the second as well because it is so similar to other poems within Li Qingzhao's collected works. I agree with Egan on poems 3.3, 3.4, 3.5, 3.23, 3.49, and 3.64. I consider poems 3.3, 3.4, and 3.5 to be some of Li Qingzhao's most unique poems in her collection, and poem

3.23 to be one of her more typical poems. My findings on Poem 3.64 agree with Egan's classification since it is historically not credible and it has such low similarity.

Of course there are many limitations within this experiment. Because the sample size is so small, only 66 poems, we were limited to a computational technique that only accounts for diction. We are unable to use the computer to make any judgments on style aside from the use of certain words. The order of the words is not taken into account, and other stylistic elements are ignored. However, we are able to comprehensively analyze the use of diction throughout the poems and use that to help our search for authorship. Through this experiment, we were able to isolate poems that displayed relatively high or low levels of similarity when compared to the rest of the anthology. We also created software that reads in a poem and outputs that poem along with the anthology's term frequencies. This is a helpful software for reading poems and highlighting common words/phrases, unique imagery, and deciding which terms are common within a poet's vocabulary. The software also is able to find all the common words between two poems and print which poems contain a certain word. These functions were very helpful in the analysis of these poems, leading to the question of what is the role of computational science within reading poetry?

The computer's algorithm works by turning a poem into a vector, a mathematical representation of data, and comparing its components. This method is completely mathematical and does not account for any of the non-statistical aspects of the diction, style, mood, synonyms, etc. Computational methods are used best in conjunction with a skilled reader's eye. It is not enough to put these poems into an algorithm and judge the authenticity of them immediately. Instead, we can use this algorithm to help us tell the poems' authenticity, the statistics behind its

diction, and how it relates to other poems. Interpreting these results, in context with each poem's style and history is crucial to properly determining authorship.

As computational techniques improve, the applications within linguistics and Chinese studies become broader and allow us to look at poetry from an alternative perspective. With this, the question arises, how should computational and traditional methods of reading poetry interact? Are we able to learn something from these computational techniques? What are some areas that computer algorithms can supplement while reading poetry? What is the computer program unable to describe and what is it's limitations? These questions are complex, but we began to understand the context of computational humanities within this thesis.

In the future, if Jieba or other natural language processing modules within certain programming languages include tokenization techniques for classical Chinese or pre-modern Chinese then computational humanities for Song dynasty poetry may become more prevalent. More research encourages new ways of thinking about text, language, and the interactions between the human reader and the computer's algorithms.

Aside from the computational side of this thesis, we also discuss authorship. The authorship behind Li Qingzhao's anthology is very interesting because it almost certainly included some imitations and misattributions. These imitations must contain enough similar diction, style, and contextual details to be attributed to her. As time passes, these misattributed poems and their emotions, styles, and stories become part of Li Qingzhao's literary legacy. The overall style, emotions, and stories within Li Qingzhao's literary footprint are shaped not only by her most authentic works, but also by her imitations. Perhaps some imitation poems embody her style and emotion so well, that they are considered to be some of the better poems by Li

Qingzhao. There is no way to definitively tell which poems are hers and which ones are not, but we can make claims about which poems seem similar enough to her style to pass.

Works Cited

Anwar, Waheed, et al. "Design and Implementation of a Machine Learning-Based Authorship Identification Model." *Scientific Programming*, vol. 2019, Jan 2019, pp 14, https://doi.org/10.1155/2019/9431073.

"Authorship Attribution Using Machine Learning." Umairacheema, https://github.com/umairacheema/authorship-attribution.

"Authorship Attribution with Python." *AICBT,* http://www.aicbt.com/authorship-attribution/. Accessed on 18 Sep. 2020.

Chang, Kang-i Sun, and Haun Saussy. *Women Writers of Traditional China: An Anthology of Poetry and Criticism*. Stanford: Stanford University Press, 1999.

Chaudhary, Varun. "Cosine similarity: How does it measure the similarity, Maths behind and usage in Python" Towards Data Science. https://towardsdatascience.com/cosine-similarity-how-does-it-measure-the-similarity-maths-behind-and-usage-in-python-50ad30aad7db. Accessed on 3 May 2021.

Chou, Ying-Hsiung. *The Chinese Text: Studies in Comparative Literature*. The Chinese University Press, 1986.

Da Pang Canjun, 答龐參軍. In Reply to Aide Pang. Poem by Tao Yuanming, 陶淵明. Translation by Tsung-cheng Lin , in How to Read Chinese Poetry in Context: Poetic Culture from Antiquity through the Tang, Columbia University Press, New York, 134, 2018.

Dan, Yao, et al. *Chinese Literature*. Cambridge: Cambridge University Press, 2012.

"Determining the Author of a Text". Wolfram Language, https://www.wolfram.com/language/gallery/determine-the-author-of-a-text/. Accessed 3 May 2021.

"蝶恋花·槛菊愁烟兰泣露." https://baike.baidu.com/item/%E8%9D%B6%E6%81%8B%E8%8A%B1%C2%B7%E6%A7%9B%E8%8F%8A%E6%84%81%E7%83%9F%E5%85%B0%E6%B3%A3%E9%9C%B2. Accessed on 3 May 2021.

Gupta, Sanket. "Overview of Text Similarity Metrics in Python." Sanket Gupta, https://towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50. Accessed on 3 May 2021.

Huang, Xuedong, et al. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. New Jersey: Prentice Hall PTR, 2001.

Huilgol, Purva. "Quick Introduction to Bag-of-Words (BoW) and TF-IDF for Creating Features from Text. " Analytics Vidhya, https://towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50. Accessed on 3 May 2021.

Idema, Wilt, and Lloyd Haft. *A Guide to Chinese Literature*. Amsterdam: Amsterdam University Press, 1996.

Iyer, Rahul, and Carolyn Rose. "A Machine Learning Framework for Authorship Identification From Texts." Carnegie Mellon University, Working Paper, Dec. 2019, https://arxiv.org/pdf/1912.10204.pdf.

Juola, Patrick. *Authorship Attribution*. Boston: now Publishers, 2008.

Koppel, M., Schler, J. & Argamon, S. Authorship attribution in the wild. *Lang Resources & Evaluation* 45, 83–94 (2011). https://doi.org/10.1007/s10579-009-9111-2

Li, Jenny, et al. “A Comparison of Classifiers and Features for Authorship Identification of Social Networking Messages.” Concurrency and Computation: Practice and Experience, vol. 29, no. 14, Jul. 2017, pp. 1-15, https://doi.org/10.1002/cpe.3918.

Lin, Shuen-fu, et al. *Senses of the City: Perceptions of Hangzhou and Southern Song China.* Hong Kong: Chinese University of Hong Kong Press, 2017.

Muller, A. Charles. “Chapter One: Free and Easy Wandering 莊子 逍遙遊.” http://www.acmuller.net/con-dao/zhuangzi.html#note--5. Accessed on 3 May 2021.

Ramnial, Hoshiladevi, et al. “Authorship Attribution Using Stylometry and Machine Learning Techniques.” *Intelligent Systems Technologies and Applications*, edited by Shireen Panchoo, 2016, pp. 113-125.

Egan, Ronald, et al. *The Works of Li Qingzhao.* Berlin: De Gruyter, 2019.

Egan, Ronald. *The Burden of Female Talent*. Boston: Harvard University Art Center. 2013.

Shou-Yi, Ch’en. *Chinese Literature: A Historical Introduction*. New York: The Ronald Press Company, 1961.

Widmer, Ellen, and Kang-i Sun Chang. *Writing Women in Late Imperial China*. Stanford: Stanford University Press, 1997.

Yang, Xiaoneng. *Reflections of Early China: Decor, Pictographs, and Pictorial Inscriptions*.Seattle: University of Washington Press, 2000.

Zhai, ChengXiang, and Sean Massung. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Morgan & Clay Publishers. 2016.

Zheng, Rong, et al. “A Framework for Authorship Identification of Online Messages:

Writing-Style Features and Classification Techniques." Journal of the American Society for Information Science and Technology, vol. 57, no. 3, Feb. 2006, pp. 378-398, https://doi.org/10.1002/asi.20316.